

DOCUMENT RESUME

ED 422 353

TM 028 905

AUTHOR Millman, Jason, Ed.; Darling-Hammond, Linda, Ed.
 TITLE The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers.
 ISBN ISBN-0-8039-4523-X
 PUB DATE 1990-00-00
 NOTE 441p.
 AVAILABLE FROM Corwin Press, Inc., A Sage Publications Company, 2455 Teller Road, Thousand Oaks, CA 91320 (paperback: ISBN-0-8039-4523-X, \$29.95).
 PUB TYPE Books (010) -- Collected Works - General (020) -- Guides - Non-Classroom (055)
 EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
 DESCRIPTORS Accountability; Educational Assessment; Educational Improvement; *Elementary School Teachers; Elementary Secondary Education; *Evaluation Methods; Evaluation Utilization; *Professional Development; *Secondary School Teachers; Self Evaluation (Individuals); Standards; Teacher Certification; *Teacher Evaluation; Test Use

ABSTRACT

The 25 chapters in this handbook examine evaluation purposes, processes, and methods, and discuss how they shape the implementation and outcomes of evaluation. Following an introduction in chapter 1, the nine chapters in Part I consider the purposes of evaluation. Chapters 3, 4, and 5 are concerned with the evaluation processes that determine selection into teaching: preservice evaluation, licensure, and certification. Chapters 6, 7, and 8 focus on how evaluation may support the development of greater teaching competencies for beginning and veteran teachers. Chapters 9 and 10 consider school accountability and school improvement. Part II examines the various methods used to evaluate teachers, along with their strengths and limitations as applied to the many purposes of evaluation. Chapters 11 and 12 deal with two methods for evaluating inservice teachers: classroom observations and ratings of classroom performance, while chapter 13 deals with teacher self-assessment. Chapters 14 and 15 deal with less widely used methods, the use of student test scores and the use of teacher portfolios. Chapters 16 through 18 address the use of formal tests to evaluate teachers for licensure or certification, and chapter 19 discusses combining evaluation data from multiple sources for decision making. In Part III, some of the system-wide concerns that developers and users of evaluation plans must confront are discussed. Chapters explore legal, ethical, administrative, and economic aspects of teacher evaluation, and ways to improve and support teacher evaluation. The use of personnel standards to improve teacher evaluation is also discussed. Each chapter contains references. (Contains six figures and four tables.) (SLD)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED 422 353

The New Handbook of **Teacher Evaluation** Assessing Elementary and Secondary School Teachers



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL IN
MICROFICHE ONLY HAS BEEN GRANTED BY

G. Aikema

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

2B

edited by JASON MILLMAN
LINDA DARLING-HAMMOND

TM028905

The New Handbook of
**Teacher
Evaluation**

The New Handbook of

Teacher Evaluation

**Assessing Elementary and
Secondary School Teachers**

**edited by JASON MILLMAN
LINDA DARLING-HAMMOND**



CORWIN PRESS, INC.

A Sage Publications Company
Newbury Park, California

Copyright © 1990 by Sage Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Distributed by Corwin Press, Inc., A Sage Publications Company

Address inquiries and orders to:



Corwin Press, Inc.
P.O. Box 2526
Newbury Park, California 91319

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Main entry under title:

The New handbook of teacher evaluation: assessing elementary and secondary school teachers / edited by Jason Millman and Linda Darling-Hammond.
p. cm.

Includes bibliographical references.

ISBN 0-8039-3394-0. — ISBN 0-8039-4523-X (pbk.)

1. Teachers—United States—Rating of — Handbooks, manuals, etc. 2. Elementary school teachers—United States—Rating of — Handbooks, manuals, etc. 3. High school teachers—United States —Rating of — Handbooks, manuals, etc.

I. Millman, Jason. II. Darling-Hammond, Linda, 1951—

LB2838.N4215 1990

371.1'44—dc20

93 94 95 96 10 9 8 7 6 5 4 3

Contents

About the Authors	7
Preface	15
1. Teacher Evaluation in Transition: Emerging Roles and Evolving Methods LINDA DARLING-HAMMOND	17
Part I PURPOSES OF EVALUATION	
2. Intended and Unintended Consequences: Purposes and Effects of Teacher Evaluation GARY NATRIELLO	35
3. Preservice Evaluation of Teachers WALT HANEY	46
4. Licensure and Certification of Teachers: An Appraisal GARY SYKES	62
5. Teacher Selection MICHAEL SCRIVEN	76
6. Assistance and Assessment for Beginning Teachers KENNETH D. PETERSON	104
7. Beyond Minimum Competence: Evaluation for Professional Development DANIEL L. DUKE AND RICHARD J. STIGGINS	116
8. Evaluating Teachers for Career Awards and Merit Pay SAMUEL B. BACHARACH, SHARON C. CONLEY AND JOSEPH B. SHEDD	133
9. Evaluation for Tenure and Dismissal EDWIN M. BRIDGES	147
10. Teacher Evaluation for School Improvement EDWARD F. IWANICKI	158
Part II METHODS OF EVALUATION	
11. Classroom Observation SUSAN S. STODOLSKY	175
12. Teacher Ratings: A Call for Teacher Control and Self-Evaluation THOMAS L. GOOD AND CATHERINE MULRYAN	191
13. Self-Assessment LARRY W. BARBER	216

14. Using Student Test Scores to Evaluate Teachers	229
GENE V. GLASS	
15. The Schoolteacher's Portfolio: An Essay on Possibilities	241
TOM BIRD	
16. Conventional Tests for Licensure	257
GEORGE MADAUS AND WILLIAM A. MEHRENS	
17. Performance Tests, Simulations, and Other Methods	278
EDWARD H. HAERTEL	
18. Setting Standards on Teacher Certification Tests	295
RICHARD M. JAEGER	
19. Combining Evaluation Data from Multiple Sources	322
WILLIAM A. MEHRENS	
Part III CROSS-CUTTING PERSPECTIVES	
20. Legal Issues Concerning Teacher Evaluation	337
MICHAEL A. REBELL	
21. The Ethics of Educational Evaluation	356
KENNETH A. STRIKE	
22. Governance Issues in the Evaluation of Elementary and Secondary School Teachers	374
ARTHUR E. WISE AND TAMAR GENDLER	
23. Economic Aspects of Teacher Evaluation	390
STEPHEN A. HOENACK AND DAVID H. MONK	
24. Embracing Contraries: Implementing and Sustaining Teacher Evaluation	403
MILBREY WALLIN MCLAUGHLIN	
25. Using the Personnel Evaluation Standards to Improve Teacher Evaluation	416
DANIEL L. STUFFLEBEAM AND JAMES R. SANDERS	
Index	429

About the Authors

SAMUEL B. BACHARACH is Professor of Organizational Behavior at Cornell University. He is also currently a senior research consultant with Organizational Analysis and Practice, Inc., Ithaca, New York. Among his books are *Power and Politics in Organizations*, *Bargaining Power Tactics and Outcomes*, *Paying for Better Teaching: Merit Pay and Its Alternatives*, and an edited volume titled *Organizational Behavior in Schools and School Districts*. He is also working with Sharon C. Conley and Joseph B. Shedd on a volume on education reform and is the author of the soon to be published book *Organizational Behavior: The Science of Analysis and the Art of Synthesis*.

LARRY W. BARBER, Director, Center on Evaluation, Development, and Research, Phi Delta Kappa, was the Assistant Superintendent for Research, Evaluation, and Development in Louisville, Kentucky (1969-74), and Eugene, Oregon, (1974-82); past vice-president of AERA (1975-77); author of *Improving Teacher Performance: Formative Evaluation* and "Teacher Evaluation and Merit Pay;" co-editor of *Improving Your Own Instruction: Self-Assessment and Peer Review*; coauthor of *Developing Research Skills for Professional Educators*; consultant to five state departments of education to establish state teacher evaluation systems; and is a professional fly fishing guide in Montana each summer.

TOM BIRD is a graduate student at Stanford University's School of Education, where he is writing a dissertation concerning pragmatic inquiries into school-teaching and is working in the School's Teacher Assessment Project. Previously, he led a study of the California Mentor Teacher Program for the Far West Laboratory for Educational Research and Development. With Judith Little at the Center for Action Research in Colorado, he conducted a study of the instructional leadership in eight secondary schools.

EDWIN M. BRIDGES is a Professor of Education and Director of the New Pathways Program for Prospective Principals at Stanford University. For the past seven years he has been studying the problem of teacher incompetence and what might be done to promote a quality teaching force. He is the author of *The Incompetent Teacher: The Challenge and the Response* and *Managing the Incompetent Teacher*. Professor Bridges recently returned from a lecture tour of

major universities in the People's Republic of China, where he lectured on teacher evaluation and teacher quality.

SHARON C. CONLEY is an Assistant Professor in the Department of Educational Administration at the University of Arizona. Her main research interests include organizational behavior in school districts, career development, performance appraisal, and job design. One of her recent publications is, "A Career Developmental Framework for Evaluating Teachers as Decision-Makers," coauthored with Samuel B. Bacharach and Joseph B. Shedd in the *Journal of Personnel Evaluation in Education* (Volume 1, 1987).

LINDA DARLING-HAMMOND was Senior Social Scientist at The RAND Corporation and Director of RAND's Education and Human Resources Program while completing work on this volume. She is currently Professor of Curriculum and Teaching at Teachers College, Columbia University. Her recent work has included studies of teacher evaluation and selection practices, teacher supply and demand, teacher testing policies, and the development of educational indicators. She received the American Education Research Association's Research Review Award in 1985 for "Teacher Evaluation in the Organizational Context: A Review of the Literature." She is coauthor of *The Evolution of Teacher Policy*, *Teacher Evaluation: A Study of Effective Practices*, *Effective Teacher Selection Practices*, and *Licensing Teachers*.

DANIEL L. DUKE is Chair of the Department of Educational Leadership and Policy Studies, Curry School of Education, University of Virginia. He has served on the faculties of Stanford University and Lewis and Clark College, Portland, Oregon. His areas of specialization are teacher evaluation as it relates to professional development and organizational theory and leadership. Dr. Duke has authored 15 books and over 100 articles on these and other topics.

TAMAR GENDLER is an education policy analyst at The RAND Corporation. She has done research and writing on questions surrounding teacher professionalism for the Council for Basic Education and the American Federation of Teachers. She is currently involved in a study of school climate in New York City and in a project to reform Minnesota's process of teacher induction.

GENE V. GLASS is Professor in the Divisions of Leadership and Policy Studies and Psychology in Education at the College of Education, Arizona State University. Dr. Glass has pursued research in measurement, statistics, and policy studies for the past 25 years. He is the author of *Meta-Analysis in Social Research* (1981) and *School Class Size* (1982), both published by SAGE.

THOMAS L. GOOD is Professor in the Department of Curriculum and Instruction and Research Associate at the Center for Research in Social Behavior at the

University of Missouri-Columbia. He is the editor of the *Elementary School Journal* and an advisory editor for five other journals. His current research interests include the role of teacher and student expectations in classroom learning and the impact of small-group instructional models on student thinking and student participation rates. His published works include *Looking in Classrooms* (with Jere Brophy) published by Harper & Row (1987) and *Educational Psychology: A Realistic Approach* (also with Jere Brophy), Longman, White Plains, New York (1986).

EDWARD H. HAERTEL is an Associate Professor in the Stanford University School of Education, where he is affiliated with the Committee on Psychological Studies in Education and the Committee on Research and Evaluation Methods. For the past two years, he has worked closely with the Teacher Assessment Project on new forms of teacher examinations, and he has also worked at the National Center for Education Statistics in Washington, DC, as a Fellow of the American Statistical Association.

WALT HANEY is an Associate Professor at Boston College, where he is also a Senior Research Associate in the Center for the Study of Testing, Evaluation, and Educational Policy, and director of the Educational Technology Program. Haney is former editor of the NCME journal *Educational Measurement: Issues and Practice*.

STEPHEN A. HOENACK is presently working at the World Bank, on leave from the University of Minnesota, where he is a Professor in the Humphrey Institute of Public Affairs and serves as director of the Management Information Division. His research interests include education, organizations, and economic development. He has published extensively on the economics of higher education and is the author of *Economic Behavior Within Organizations*. He received his Ph.D. in Economics from the University of California at Berkeley.

EDWARD F. IWANICKI is Professor of Educational Leadership at the University of Connecticut. He is responsible for teaching courses in the areas of supervision as well as teacher and program evaluation. He has worked extensively with school systems and state agencies on policy as well as procedural issues in the design, implementation, and evaluation of teacher evaluation programs. He is editor of the new *Journal of Personnel Evaluation in Education*, and author of a forthcoming book, *Teacher Evaluation for Professional Growth and Accountability*.

RICHARD M. JAEGER is Professor of Educational Research Methodology and Director of the Center for Educational Research and Evaluation at the University of North Carolina at Greensboro. He is author of *Sampling in Education and the Social Sciences* (1984) and *Statistics: A Spectator Sport* (1983),

editor of *Complementary Methods for Research in Education* (1988), and coeditor of *Minimum Competency Achievement Testing* (1980). He is past President of the National Council on Measurement in Education and past editor of the *Journal of Educational Measurement*. He currently serves as series editor of *Measurement Methods for the Social Sciences* and is on the editorial boards of several measurement journals. His research interests include competency testing, standard-setting methodology, and certification testing.

GEORGE MADAUS is Professor of Education and Director of the Center for the Study of Testing, Evaluation, and Educational Policy at Boston College. He is also Executive Director of the National Commission on Testing and Public Policy. Professor Madaus has written extensively in the area of teacher testing and has served as an expert witness in several teacher testing lawsuits.

MILBREY WALLIN MCLAUGHLIN is an Associate Professor of Education and Director of the Center for Research on the Context of Secondary School Teaching at Stanford University's School of Education. Prior to joining Stanford, she was a social scientist with The RAND Corporation. McLaughlin's areas of research specialization include planned change in education, policy implementation, and the secondary school workplace. Her present work focuses on the factors that comprise the context of secondary school teaching and the ways in which they shape conceptions of teaching in diverse settings.

WILLIAM A. MEHRENS is a distinguished professor of measurement at Michigan State University. He is past president of both the National Council of Measurement in Education (NCME) and the Association for Measurement and Evaluation in Guidance (AMEG) (currently called the Association for Measurement and Evaluation in Counseling and Development). He is a coauthor of several text books in measurement and educational psychology and coeditor of several books of readings. He is widely published in various measurement journals. His more recent publications have frequently dealt with issues in teacher licensure tests. He is a consultant to several state departments of education, advising them on both student and teacher assessment activities.

JASON MILLMAN is Professor of Educational Research Methodology at Cornell University. Among other publications, he edited the *Handbook of Teacher Evaluation* (Sage, 1981), the *Educational Researcher*, and the *Journal of Educational Measurement*. Offices held include President of the National Council on Measurement in Education, Vice-President of Division D (Measurement and Research Methodology) of the American Educational Research Association (AERA), and Member-at-Large of AERA. He has consulted extensively on technical questions related to educational testing generally and teacher evaluation specifically.

DAVID H. MONK is Associate Professor of Educational Administration at Cornell University. He earned his doctorate at the University of Chicago and has taught at the University of Rochester, the International Institute for Educational Planning in Paris, and the Institut de Recherche sur l'économie de l'éducation at the University of Burgundy in Dijon, France. In addition to numerous articles dealing with economic aspects of educational administration, he is the senior editor of *Microlevel School Finance: Issues and Implications for Policy* (Ballinger) and the author of *Educational Finance: An Economic Approach* (Random House).

CATHERINE M. MULRYAN is a Lecturer in Education at St. Patrick's College of Education, Dublin, Ireland, and is at present working on her doctoral degree in Curriculum Instruction at the University of Missouri-Columbia, where she also works as a research assistant with Dr. Tom Good — who also advises her doctoral work — at the Center for Research in Social Behavior. She received her B.Ed. degree in Elementary Education at St. Patrick's College, Dublin, and her M.Ed. at the University of Dublin, Trinity College. Her main academic interests are in the classroom research area, and her main research interest at present is in the area of student mediation of instruction in the context of cooperative small-group teaching. She has published in *Irish Educational Studies*.

GARY NATRIELLO is Associate Professor of Sociology and Education in the Department of Philosophy and the Social Sciences at Teacher's College, Columbia University. His research interests include processes for the evaluation of teachers and students and their impacts, the structure of schools and classrooms, and the education of disadvantaged students. Recent publications have included "The Impact of Evaluation Processes on Students," which appeared in *Educational Psychologist* in 1987, and "The Impact of Evaluation Processes on Performers: The Case of Teachers," which appeared in *Advances in Educational Administration* in 1988. Professor Natriello is currently involved in a longitudinal study of the impact of teacher education and workplace conditions on teachers' beliefs and attitudes.

KENNETH D. PETERSON is an Associate Professor at Portland State University. His research and teaching specialties are teacher education and teacher evaluation. Dr. Peterson has been a principal investigator in a series of United States Department of Education research grants to study teacher evaluation.

MICHAEL A. REBELL is the founding partner of Rebell & Katzive, a New York City law firm that specializes in education law. He is also a Visiting Lecturer at Yale Law School. Rebell has litigated a number of major class action litigations involving testing issues and currently serves as Special Master in *Allen v. McDonough*, the Boston special education litigation. He is coauthor of *Educational Policy Making and the Courts* (University of Chicago), *Education*

and *Equality* (Princeton University), and has written numerous articles and monographs on subjects involving testing, law, and education.

JAMES R. SANDERS is Professor of Education and Associate Director of The Evaluation Center at Western Michigan University. He received his master's degree in Educational Research from Bucknell University and his Ph.D. in Educational Research and Evaluation from the University of Colorado. He is coauthor of *Educational Evaluation: Theory and Practice*, *Practices and Problems in Competency-Based Measurement*, and *Educational Evaluation: Alternative Approaches and Practical Guidelines*, and he is author of many articles, monographs, and technical reports in the areas of instructional research and research and evaluation methodology. In 1988, he was elected Chairman of the Joint Committee on Standards for Education Evaluation.

MICHAEL SCRIVEN took his first degrees in Mathematics at the University of Melbourne, followed by a doctorate in Philosophy at Oxford. His first publications were in psychology and computer simulation, with later ones in logic and the philosophy of science, particularly explanation theory. Beginning with some years of work on program and product evaluation, he became increasingly involved with educational research when at Berkeley, and has recently been running the Centre for Tertiary Education Studies at the University of Western Australia, publishing on teacher evaluation and the philosophy of technology, and putting out two newsletters on microcomputer studies.

JOSEPH B. SHEDD is a senior consultant with Organizational Analysis and Practice, Inc., of Ithaca, New York. His most recent publications include, "Collective Bargaining, School Reform, and the Management of School Systems," (*Educational Administration Quarterly*, November 1988); "Teacher Participation in the Management of School Systems," coauthored with Sharon C. Conley and Timothy Schmidle (*Teachers College Record*, Winter 1988); and "A Career Developmental Framework for Evaluating Teachers as Decision Makers," coauthored with Samuel B. Bacharach and Sharon C. Conley (*Journal of Personnel Evaluation in Education*, September 1987). He is presently principal consultant to the Cooperative Relationships Project, a New Jersey State Education Department program to enhance cooperation and promote teacher empowerment in New Jersey's public schools.

RICHARD J. STIGGINS is Director of the Center for Classroom Assessment, Northwest Regional Education Laboratory, Portland, Oregon. He is former director of test development for the American College Testing Program and has taught at Michigan State University, University of Minnesota, and Lewis and Clark College. Dr. Stiggins specializes in performance assessment methodology and classroom assessment.

SUSAN S. STODOLSKY is Professor of Education and Psychology, University of Chicago. Professor Stodolsky has interests in classroom instruction, the development of student interests and attitudes, and evaluation methodology. She has paid particular attention to subject matter as an influence on educational practice and participants' perceptions. Her 1988 book, *The Subject Matters: Classroom Activity in Math and Social Studies*, was published by the University of Chicago Press. A relevant article, "Teacher Evaluation: The Limits of Looking," appeared in *Educational Researcher*, November 1984.

KENNETH A. STRIKE is Professor of Philosophy of Education at Cornell University. He has done research on issues of ethics, political philosophy, and educational policy. Recent publications include *Liberal Justice and the Marxist Critique of Education* (forthcoming, Routledge), *The Ethics of Teaching* with Jonas Soltis (Teachers College Press), and *The Ethics of School Administration* with Jonas Soltis and Emil Haller (Teachers College Press). He is also a devotee of canoeing, a preoccupation which, while unrelated to teacher evaluation, demonstrates breadth of interest, strength of judgment, soundness of character, and fitness of body.

DANIEL L. STUFFLEBEAM is Professor of Education and Director of The Evaluation Center, Western Michigan University. He participated in the emergence of the program evaluation discipline in the 1960s. He developed the CIPP Model, one of the first and still most widely used evaluation approaches; helped establish evaluation departments in many local, state, and national agencies; organized the well-known Evaluation Center; and authored or coauthored 10 books and about 70 articles and book chapters. From 1975-1988, he chaired the Joint Committee that issued national standards for both program and personnel evaluations. For theory contributions, he received the 1984 Western Michigan University Distinguished Faculty Scholar Award and the 1985 American Evaluation Association Paul Lazarsfeld prize. He is coeditor and frequent contributor to Kluwer Publishers' series of books on evaluation in education and human services.

GARY SYKES is an Assistant Professor in the Department of Educational Administration, College of Education, at Michigan State University. His interests include policy analysis and the study of teaching and school reform. He is currently working with both the Holmes Group and the National Board for Professional Teaching Standards.

ARTHUR E. WISE is director of The RAND Corporation's Center for the Study of the Teaching Profession in Washington, DC. The Center conducts research and advises policymakers on issues related to teacher professionalism. His 1979 book, *Legislated Learning*, which preceded the current wave of reform, called for teacher professionalism, school-based management, and local control

of the schools. As a chief consultant with commissions in Connecticut and New York, he helped to develop major new legislative proposals to professionalize teaching.

Preface

The word *New* in the title of this handbook is descriptive. This handbook is very different from the 1981 edition because the times are different, the authors are different, and the contents are different.

In 1981, much attention to teacher evaluation was at the college level where student ratings of instructors were the focus of most research. At the elementary and secondary school level, most teacher evaluation decisions resulted from classroom observations more or less conducted as they had been for decades previously.

In 1981, only seven states had implemented a teacher testing program for certification; now almost all states have such a program. Since 1981, the National Teacher Examinations (NTE) have undergone a major revision, and an even more substantial revision is being planned for the future. A National Board for Professional Teaching Standards has been established to consider innovative methods of board certification for outstanding teachers, and career ladder plans to reward such teachers have been erected at a limited number of sites. Meanwhile, many states have enacted guidelines or requirements for on-the-job evaluation of beginning teachers — and sometimes of veteran teachers as well.

Since 1981, other events signifying increased interest in teaching evaluation are evident. A new periodical, the *Journal of Personnel Evaluation in Education*, was begun; a special issue of *Educational Measurement: Issues and Practices* was devoted to teacher evaluation; and the number of teacher-evaluation sessions at the annual meeting of the American Educational Research Association has risen from two to seven. National studies of teacher-evaluation practices have been commissioned. We know much more about good and poor practices and helpful and less helpful institutional and social contexts than we did when the first volume was published.

Having recently conducted a review of evaluation practices in the largest 100 school districts, one researcher bemoans the finding that current practices do not reflect state-of-the-art knowledge. One purpose of this volume is to help close this gap.

With few exceptions, the authors of this new edition are different from those of the 1981 work, and the holdovers are, by and large, addressing different issues than those addressed previously. The content of the 25 chapters of the present handbook has little overlap with the 17 chapters of its predecessor. Even in the

few cases where the chapter titles promise similar coverage, the treatment is different — in part because we have a greater understanding of the issues and a greater base of experience than before.

A principal difference in content between the two editions is that the present work is limited to teacher evaluation at the elementary and secondary level. It is at this level of schooling in which most of the changes have occurred and because college faculty evaluation was seen to involve very different issues, better treated in a separate volume.

Similarities with the 1981 volume remain. The National Council for Measurement in Education (NCME) agreed to sponsor this volume as they had the previous one. All royalties from the sale of this book, as before, go to support the work of NCME. Sage Publications continues to be the publisher for the *Handbook*.

We are indebted to Shirley Preston at Cornell and Nancey Rizer at the RAND Corporation, who were responsible for the day-to-day management of correspondence. The authors benefited greatly from a large number of colleagues who served as their advisors. Their contributions are recognized in the initial footnotes of the individual chapters. The reader will benefit most from the fact that we were fortunate to recruit an outstanding collection of authors, who are among the most visible and talented scholars in the field of teacher evaluation.

Jason Millman

Linda Darling-Hammond

1. Teacher Evaluation in Transition: Emerging Roles and Evolving Methods

Linda Darling-Hammond

This *Handbook* is being published at a time when the role of evaluation in schools and in the teaching occupation is being reshaped in important ways. This reshaping is a result of an increased focus on teachers in the policy environment, of the growing sophistication of basic and applied research on teaching and teacher evaluation, and of willingness of practitioners to engage many of the difficult issues that evaluation poses.

Two major concepts characterize the educational reform movement that has gathered momentum throughout the 1980s and promises to extend well into the 1990s. *Teacher professionalism* and *school restructuring* are the major watchwords used to describe efforts to reform teaching and schooling so that they will focus more directly on learners' needs. The many initiatives launched during this reform movement seek to improve education by recruiting, preparing, and retaining competent teachers and by better utilizing their knowledge and talents over the course of a reshaped career. These initiatives rely on teacher evaluation for a wide variety of purposes, including selection, training, improvement, and advancement. They often envision broader roles for teachers in evaluation and for evaluation in schools. As schools are asked to define their own improvement strategies, agendas for individual evaluation and organizational renewal are increasingly intertwined.

In the past, teacher evaluation has generally not been a high-stakes activity, in part because improving the quality of teachers has not been seen as critical for improving the quality of education. Instead, school improvement efforts over the past several decades have focused on improving the curriculum, altering school management methods, and developing new programs. Thus, teacher evaluation, where practiced, was often an exercise to which few resources and little organizational attention were devoted. As a consequence, teacher evaluation has often had little influence on decisions about personnel, staff development, or the structure of teaching. As more attention is being devoted to evaluation, and as its results are used for a greater range of decisions, its role in shaping teaching will increase. Educators must, therefore, consider *how* evaluation affects teaching performance, rather than whether or not evaluation *will* affect it.

The Policy Context

The public has come to believe that the key to educational improvement lies as much in upgrading the quality of teachers as in revamping school programs and curricula. Foreshadowing the reforms of the 1980s, the most frequent response to the 1979 Gallup poll's question on what public schools could do to earn an "A" grade was improving teacher quality, beating by large margins such reforms as emphasizing the basics, improving school management, lowering class size, or updating the curriculum. Importantly, those other approaches to reform, which often hypothesized a teacher-proof road to educational improvement, had characterized state legislative initiatives throughout the 1970s. In response to these new perceptions about the importance of teacher quality, states and local school districts have initiated a wide range of policy changes affecting the certification, evaluation, and tenure of both prospective and currently employed teachers.

At least 46 states have adopted teacher-competency tests, such as the National Teacher Examinations, as a prerequisite for teacher certification; 25 have required tests for admission to teacher-education programs. Most states have replaced lifetime teaching certificates with requirements for continuing licensure. Some have adopted comprehensive programs that include higher admission standards for colleges of education, competency tests for certification and recertification, evaluation of performance, and continuing teacher education.

Most states have legislated requirements for teacher-performance evaluation and some of the more recent statutes specify which testing instruments or evaluation procedures are acceptable. Increasingly popular are state-mandated, beginning-teacher programs that prescribe the supervision and evaluation process for first-year teachers. When used as a licensing activity, these beginning-teacher programs are presumably distinct from evaluation for employment decisions; however, they are inextricably entangled with local district procedures for assessing teaching performance. Many states have also mandated the use of these or other procedures for merit pay or career ladder placement determinations.

Clearly, the development of teacher-evaluation practices in local school districts does not occur in a vacuum. State policies often define some of the key features of evaluation; other state and local policies regarding teachers and teaching define the nature of teaching desired and the means by which it is sought. These means include everything from the job roles and tasks assigned to teachers to teacher selection and assignment policies to instructional management systems.

Evaluation in Organizations: Intentions and Implementation

Teacher evaluation can be a routine, pro forma activity with little utility for shaping what goes on in schools, or it can be an important vehicle for communicating organizational and professional norms and for stimulating improvement. This Handbook starts from the proposition that the outcomes of evaluation often depend as much on the conditions under which it is designed and implemented as on the formal design as it exists on paper. Evaluation processes, their outcomes and effects, are a function of many different technical, organizational, and political factors that interact in important ways.

Technical aspects of evaluation include (a) methods, instrumentation, and sources of evidence; (b) the training and expertise of evaluators; and (c) structural features of the evaluation process, such as who evaluates, when and how often, how data are combined and aggregated, what purposes evaluation is intended to serve, how judgments are communicated, and what follow-up is planned. A fair amount of research attention and practitioner energies is devoted to designing and implementing singular technical features of evaluation. Until recently, less attention has been paid to the combined outcomes of their interaction as a total system of evaluation. Each of these factors is shaped, in turn, by organizational conditions and constraints.

Organizational factors influencing evaluation include school or school district goals and perceived problems; resources such as time, personnel, and expertise for evaluation; collective bargaining and legal requirements; and structural features of the organization, such as the degree of centralization of school functions, specialization of tasks, and the size and management approach of an organization. Evaluation practices that are highly successful in some organizations may be absolutely unmanageable in others, unless substantial changes to the organizational environment are made.

Compatibility considerations arise where technical and organizational factors meet. The extent to which an organization's purposes will be achieved by the evaluation processes chosen depends on the degree to which particular methods and instruments provide reliable and valid data for the *primary purpose(s)* for which they are used; the degree to which the process as implemented is sufficiently timely, credible, and efficient to provide usable information; and the degree to which the process supports organizational norms and conceptions of good teaching.

Issues in Implementation

The application of research-based teacher-evaluation models to real-life settings must overcome the gap that exists between technically defensible specifications of criteria or methods and politically viable solutions to organizational problems. There is a growing recognition that any kind of evaluation activity

involves value choices — and conflicts — at all levels of the operating system. Evaluation is political because it serves as a tool in a larger policy-making process and because it is inherently directed at making a judgment of worth about something. Any such judgment ultimately rearranges or reaffirms an existing constellation of stakes that individuals or groups have in what is being evaluated.

Stakeholders may have divergent views of the primary purpose of teacher evaluations and, hence, of what constitutes a successful evaluation system. Differing priorities also affect implementation, because even after a policy is adopted, its terms and emphases are renegotiated at every level in the implementation system. This renegotiation may not occur in a formal way, but practices at the school district, school, and classroom levels will be a function of cross-pressures that may alter the formal process in important ways.

Implementation of any school policy, including a teacher-evaluation policy, represents a continuous interplay among diverse policy goals, established rules and procedures (concerning both the policy in question and other aspects of the school's operations), intergroup bargaining and value choices, and the local institutional context. The local implementation process and organizational characteristics — such as institutional climate, organizational structures and incentives, local political processes, expertise, and leadership style — are critical elements in determining the ultimate success of a policy at achieving its intended effects.

All of these factors argue for understanding teacher-evaluation plans in the context of organizational policies and practices. The chapters in this Handbook examine evaluation purposes, processes, and methods, and discuss how they shape the implementation and outcomes of evaluation. Though three separate sections have been created to address evaluation purposes, methods, and cross-cutting issues, the interrelationships among technical, organizational, and social concerns are treated throughout the volume. The collection also recognizes that even as a variety of factors influence evaluation, so evaluation itself influences schools and teaching.

Evaluation Effects

Evaluation involves collecting and using information to judge the worth of something. Every teacher-evaluation system must embody a definition of the teaching task and a mechanism to evaluate the teacher. Different conceptions of teaching work imply different ways by which information is collected and judgments of worth are made. Implied in these different conceptions of teaching work are different notions of educational goals, teacher knowledge and activities, teaching behavior, and self- or other-evaluation activities.

Thus, evaluation is not only influenced by organizational considerations, it also shapes the organizational context and the conditions of teaching work. Whether intentionally or not, a teacher-evaluation system represents the incen-

tive structure and mode of accountability implicitly adopted by an organization or profession. It communicates conceptions of teaching and expectations regarding performance priorities, norms for behavior, and the nature of the work itself.

Depending on how an evaluation process is designed, and how well it is implemented, it can guide professional and personal development, and influence motivation. If designed appropriately and implemented with sufficient attention, evaluation can provide data for personnel decision making, thus shaping the composition of the teaching force. These functions of evaluation are more likely to occur if evaluation is a "high stakes" activity; that is, one that is used for decision making by the teacher, the evaluator(s), and/or the organization. This is not so much a matter of intent as it is of actual implementation. Usable evaluation is not achieved by exhortation. As these chapters demonstrate, an evaluation process must be credible, appropriate for its purposes, and doable within time, personnel, and budget constraints if its results are, in fact, to be used.

Part 1 — Purposes of Evaluation

The multiple purposes and wide-ranging effects of evaluation are elegantly described by Gary Natriello in his chapter on intended and unintended consequences. He notes that an evaluation process will have a set of explicit purposes, but that its effects may extend into other areas of organizational life than those at which it is intentionally directed. These effects may manifest at the individual level, as teachers themselves experience the evaluation process; at the organizational level, as the process affects communication, performance, and sentiments within a school; and at the environmental level, as the process and practices may affect other institutions or the broader community. The consequences take many forms: evaluation may affect teacher motivation, knowledge, satisfaction, communication, consensus, levels of trust, and community confidence, as well as decisions — by both evaluators and teachers themselves — about who enters and remains in teaching.

Natriello's chapter alerts us to the complex relationships among stakeholders and participants in the educational system as it illustrates the many ways in which evaluation impacts upon organizational and classroom life. The effects of a process at each level may be positive or negative in the context of particular goals, depending upon the strategies selected and the manner by which they are implemented. And the effects may be construed differently from diverse vantage points. "Improvement," we are reminded, "is in the eye of the beholder."

Natriello provides a framework for the remaining chapters in this section as he describes three major purposes of evaluation: (1) influencing the performance of individuals within particular positions; (2) guiding decisions about the movement of individuals into and out of positions (e.g., certification and hiring, granting tenure or firing, movement to a new office or level); (3) and legitimizing the organizational control system itself; that is, persuading employees and

constituents of the organization that personnel activities are performed fairly and effectively.

The distinction between the first two categories is an important one. Subsequent chapters illustrate how designing and managing the improvement functions of evaluation raise different considerations from those associated with personnel decision-making functions that determine who teaches and in what capacities. The third category foreshadows the range of environmental issues associated with the management, governance, and implementation of evaluation systems that are tackled in part three of this volume.

*Constructing the Pool:
Initial Stages of Teacher Selection*

Chapters 3, 4, and 5 are concerned with the evaluation processes that determine selection into teaching: preservice evaluation, licensure and certification and hiring. Walt Haney reminds us that, even while initiatives to revise admissions and exit criteria for teacher education are underway, labor market forces define the supply pool for teaching in powerful ways. He reviews general college admissions and grading practices, arguing that they affect the pool of prospective teachers at least as much as more targeted strategies do, and he concludes that recruiting greater numbers of prospective teachers may do more to improve the caliber of preservice teacher selection than overreliance on singular testing or grading mechanisms.

Gary Sykes reaffirms the importance of labor market forces in determining how selection mechanisms operate. He points out that occupational credentialing processes both generate and restrict supply as they regulate the joint fluctuation of wages and standards. The supply of prospective practitioners may determine what categories of credentials are created and how standards are set and maintained. In turn, these decisions by credentialing agencies may generate a new supply of potential practitioners. In teaching, for example, state governments have created emergency, temporary, provisional, and alternative certificates to increase the size of the supply pool at a given wage.

Credentialing processes serve other functions as well. They construct the categories of competence that will define preparation and assignment, and they invent conceptions of quality. In these regards, licensure processes — those used by states to determine who will be permitted to practice — may differ from certification processes (those controlled by members of a profession to certify special or advanced competence in a field). In teaching, the recent creation of a certifying body (the National Board for Professional Teaching Standards) will interact in as yet unknown ways with existing state boards and licensing agencies. Sykes' chapter helps us to consider how the interactions of licensure and certification requirements will influence teacher supply and quality, as well as conceptions of teaching competence.

It is the merger of these forces that brings a pool — or, in times of shortage, perhaps a puddle — of candidates to the door of a schoolhouse or school district personnel office. Michael Scriven walks us through the complex thicket of considerations associated with teacher hiring, pointing out the utility and limits of a range of hiring methods as well as the importance of weighing and balancing the sometimes competing dimensions of worth and merit.

The concepts of worth, merit, and value so aptly illustrated in this chapter are fundamental to all evaluative decisions. As Scriven describes them, *merit* is a measure of the extent to which an individual measures up to some accepted standard; *worth* describes the additional value of an individual to an institution arising from other qualities than those measured by the more limited standard used to assess merit. *Value* is the sum of the two. These concepts allow an explicit recognition and a systematic evaluation of the ways in which a broad variety of individual traits, capacities, and dispositions may contribute to the larger mission of a school organization or an educational system. Frequent mentions throughout this volume of the need for multiple measures and wide-ranging criteria in evaluation stem from this understanding that value in teaching is not a unidimensional construct.

Evaluation for Individual Improvement: Supporting Teacher Development

Chapters 6, 7, and 8 focus on how evaluation may support the development of greater teaching competence for beginning and veteran teachers who have made it through preservice evaluation processes, licensure, and initial hiring. Kenneth Peterson addresses issues of assistance and assessment for beginning teachers from the perspective that greater investments in evaluation and support at this juncture have pay-offs throughout a teacher's career. Daniel Duke and Richard Stiggins extend this discussion into the later stages of professional development when the achievement of minimal competence is transformed into the pursuit of professional excellence. Samuel Bacharach, Sharon Conley, and Joseph Shedd apply principles of professional development and evaluation to the issues associated with creating a staged career for teachers.

Noting that school districts have rarely treated beginning teachers differently from veterans (expect to give them less desirable teaching assignments), Peterson describes a number of specific strategies for successful teacher induction that can serve two important educational goals: (1) developing effective teachers who take initiative, are self-regulating, and contribute to the professional collective; and (2) retaining promising practitioners. Peterson accurately notes that high attrition rates in the early years of teaching have spillover effects on the composition of the teaching force. The loss of so many beginning teachers also lowers the benefit-cost ratio of ongoing recruitment, selection, and evaluation activities.

In his review of the literature, Peterson identifies the distinctive and predictable needs of beginning teachers for information, skill development, and an interpretive lens that fosters productive long-range habits of reflection and consultation as part of professional socialization. He then describes how evaluation can be structured to provide assistance, useful feedback, needed reassurance, and positive socialization experiences for the new teacher as well as adequate performance data for public and organizational purposes. Key features include proactive support, multiple data sources, feedback tied to opportunities for inservice education, and teacher involvement so that teachers become producers and consumers of evaluation data over the course of their careers.

Duke and Stiggins adopt a similar view of professionally developing teachers as producers and users of evaluative information. Among the characteristics they note as linked to productive professional growth are an openness to change and criticism, willingness to experiment, and orientation to look for possible solutions to problems, which tend to be associated in turn with high levels of personal efficacy produced by the "constructively restless" quest for ways to improve.

Evaluation mostly useful for producing adult learning, growth, and further reflection among already skilled performers is different in important respects from that oriented toward ensuring school accountability and minimal teacher adequacy. It is more dependent on the evaluator's intimate knowledge of classroom context, subject matter content, and method; his or his credibility; and a capacity to provide detailed feedback and suggestions for the teacher's particular felt needs and classroom setting, along with a persuasive rationale for their use. For "growth beyond minimum competence," this kind of evaluation also seems to require an environment which encourages high levels of candor, self-criticism, and risk-taking. These authors suggest this may require some separation of activities aimed at professional development from periodic summative evaluation exercises aimed at documenting basic adequacy.

In chapter 8, Bacharach, Conley, and Shedd also suggest that developmentally oriented evaluation should both distinguish between documenting minimal competence and encouraging higher levels of competence, in the first instance, and between formative and summative evaluation events in the second. Arguing that while standards of minimum competence are necessary, they are insufficient to foster continuing improvement, the authors posit a three-stage pattern of career development with corresponding levels of evaluation. These developmental standards would follow a teacher from the initial stage when she or he learns to apply basic skills in routine situations, to the stage when she or he learns to integrate skills and to adapt to the unexpected, and finally to a stage where innovation, collaboration, and leadership are the benchmarks of evaluation.

In reviewing the lessons learned from efforts to implement merit pay and career ladder plans, the authors distinguish between competitive standards and developmental standards for evaluation. The developmental standards they suggest are intended to support increasing capacities for effective teacher decision

making and to allow for staged teacher compensation without dysfunctional competition that might undermine the organizational culture within the school.

*Evaluation for the Greater Good:
Accountability and School Improvement*

Supporting the continued learning of individual teachers is surely important, but it is not the only goal of teacher evaluation. Chapters 9 and 10 treat two other important objectives of evaluation: school accountability and school improvement. The first of these, addressed by Ed Bridges's chapter on evaluation for tenure and dismissal, is an absolutely critical dimension of evaluation — one much discussed but, as Bridges notes, too rarely handled effectively. With straightforward honesty, Bridges describes the various ways schools may tolerate inadequate teacher performance, especially during times of teacher shortage and especially when parents do not complain. He notes that the students most likely to be victimized by poor teachers, given the way "escape hatches" tend to operate, are less advantaged students: minorities, the poor, or the low achieving.

Bridges points out how crucial the tenure decision is, for it is after that point that teachers are presumed competent by both colleagues and the courts. He describes an evaluation system for tenure decisions that relies on multiple sources of evidence obtained at different points in time and in different teaching contexts, distinguished by the types of students being taught. If tenure decisions were more carefully made, the possibilities of having to dismiss a tenured teacher should become rarer. Still, this possibility cannot be eliminated. Bridges describes how evaluation plans can structure procedures and resources for assistance to tenured teachers having substantial difficulties in ways that will support improvement where possible and support personnel action if necessary.

Evaluation can also be turned to the goal of school improvement, as Ed Iwanicki describes. Particularly where goal-setting models of evaluation are used (in which teachers set individual goals for improvement and a plan to accomplish and evaluate those goals), school-wide improvement objectives can be combined with teachers' own plans to focus attention on school-wide priorities. Iwanicki discusses how school-wide goals might be set in ways that build collegiality and a sense of common purpose. He then describes how staff development and evaluation activities can be pointed toward those goals, using inquiry and reflection as a basis for collective growth.

Part 2 — Methods of Evaluation

Part two of this volume examines the various methods used to evaluate teachers, along with their strengths and limitations as applied to the many purposes of evaluation. These chapters demonstrate that both the range of methods applied to teacher evaluation and the body of research about these methods

have grown substantially in recent years. A common theme throughout the section is that any single method is most useful in combination with others that complement the data it can provide, and that the set of methods likely to be most useful may vary from one purpose or function of evaluation to another.

A second theme, as Susan Stodolsky first observes in the section's lead chapter and as subsequent authors affirm, is that "any method of evaluation is accompanied by a conception of what teaching is, and more specifically a conception of what good or effective teaching is." The criteria or categories built into any method focus attention on various aspects of teaching, emphasizing some and, necessarily, deemphasizing and excluding others. Even the choice of a method to be used will define in some fashion what is to be known about teaching, and how we will seek to know it. This recognition by the authors that "procedures are not neutral" enriches the technical discussion of methods with deliberation about the meaning of methods in the context of teaching.

Method for Evaluating On-the-Job Performance

Chapters 11 and 12 treat two time-honored methods for evaluating inservice teachers: classroom observations and ratings of classroom performance. The two, of course, often dovetail: observations may be one source of ratings data; ratings are also often incorporated into an observation event. Together, these chapters provide a comprehensive look at how observations and ratings have functioned as the cornerstones of teacher-evaluation systems, and how they might be used even more productively in the future.

Susan Stodolsky begins by pointing out that classroom observations are good for looking at some teacher behaviors or actions, but they are a very limited tool for gaining an understanding of teachers' thinking, or for assessing how teachers plan, evaluate and create instructional materials, choose their methods, work with colleagues and parents, evaluate students' needs, or offer feedback to students. Furthermore, some kinds of observation approaches and instruments are better suited than others for recording, analyzing, and understanding certain classes of teaching behaviors in a variety of teaching contexts.

Having framed the possibilities for and choices among observation methods, Stodolsky discusses the implications of research on teaching and its observation for high- and low-inferences methods, and for open and closed systems of evaluation. She notes that the stability, generalizability, reliability, and validity of measurements based on classroom observation are influenced by the variability of teaching behaviors in different contexts. Interpreting this variability and assessing the appropriateness of teacher behaviors require attention to a number of practical and procedural matters, to evaluator expertise and training, and to the ways in which observations are combined with additional sources of information about teaching.

Thomas Good and Catherine Mulryan reiterate the dependence of observations and ratings on particular conceptions of teaching. They trace the history of teacher-rating scales through various eras of research and opinion, showing how prevailing views of teaching influenced the characteristics of teaching and teachers evaluated. From lists of traits to activity reports to analyses of teacher-pupil interactions, rating scales have been developed from a variety of vantage points and used for a wide range of purposes by many types of raters (supervisors, principals, researchers, students, and teachers themselves).

Good and Mulryan provide a careful analysis of the uses, strengths, and limitations of various approaches to rating teachers, relating these approaches to research on teaching effectiveness as well as to different evaluation functions. They argue for the further development and use of descriptive feedback scales that can take into account the multidimensional nature of teaching and the importance of context. As a tool for stimulating reflection and discussion about teaching among teachers, such scales might take rating systems beyond the task of producing a summary score about teachers to become a useful vehicle for ongoing professional development as well.

Following this lead, Larry Barber's treatment of teacher self-assessment in chapter 13 also stresses the utility of methods that can help teachers gather data about their own teaching to guide self-reflection. Noting that the effectiveness of formative evaluation relies on the engagement and motivation of the individual teacher, Barber illustrates how self-assessment strategies can provide the necessary psychological frame for ongoing self-improvement, as well as the information needed to perceive events accurately and design avenues for change.

Barber describes many tools for self-assessment, ranging from peer review and self-study materials to videotape feedback, guided observations of other teachers, and observations or ratings by others (students, peers, consultants). Stressing the need for hybrid systems of evaluation that provide multiple sources of information, Barber places self-assessment in a broader context, noting that its inclusion in formative evaluation systems can awaken and launch the most powerful source of improvement: the teacher's own capacity of self-evaluation.

Chapters 14 and 15 treat teacher evaluation methods that, until relatively recently, have been less widely used — the use of student test scores for evaluating teachers and the use of teacher portfolios. Though very different, each of them seeks to fill in a perceived gap in the sources of data about teaching used in traditional evaluation systems. They seek to reveal, at least in part, what has happened as a result of teaching.

Gene Glass treats a range of complex questions about the use of student test scores in teacher evaluation by presenting a synthetic account of how a wide range of measurement considerations might actually play out in a system that tightly tied evaluations to student scores. Taking account of the practical and political factors that influence testing in real world settings, he creates a provocative scenario that illustrates the motivational possibilities and pitfalls of such use, as well as the potential effects of teaching and learning.

Glass's tale also illuminates the more general phenomenon of "high stakes" measurement. He demonstrates how the use of any measure to allocate rewards or resources may influence the validity of the measure itself, along with school priorities. This general lesson applies to a variety of evaluation methods and measures, and might be applied in across a wide range of evaluation contexts and decisions.

An alternative form of gathering information about the outcomes of teaching is the portfolio, now used in a few teacher evaluation systems (e.g., Tennessee's career ladder program and Georgia's beginning-teacher program), and being considered for inclusion in others, such as the system being created for the National Board for Professional Teaching Standards. Tom Bird's essay on "The Schoolteacher's Portfolio" examines the use of such means for aggregating information about work in other occupations, and discusses how teaching portfolios might be constructed and used.

Bird reminds us that useful records of work exist not only in the formal portfolios of architects and artists but also in the logs and files of pilots, lawyers, and social workers, as well as the sample cases of salespersons and distributors. The differences have to do with both the nature of the work and the use of the records as either work tools or displays for judgment. The different purposes will relate to different uses of portfolios in evaluation. In teaching, portfolios might include plans, materials, assignments, tests, and feedback to students, along with videotapes, evaluations, transcripts, and written testimonials from parents, students, or others. Noting both the promise and limitations of various kinds of evidence for different portfolio uses, Bird offers insights that sort through the tangled webs of evaluation functions in schools and the teaching profession.

Methods for Testing Teachers

Chapters 16 through 18 address an area of increasing prevalence and concern: the use of formal tests to evaluate teachers for licensure or certification. State licensure tests have become more widespread over the past decade, while the creation of a National Board for Professional Teaching Standards is intended to launch a professionally controlled certification-testing process as well. As the use of tests to evaluate teacher knowledge and skills has become more widely accepted, attention has begun to turn from earlier political concerns toward more substantive concerns for using testing methods to their best advantage.

George Madaus and William Mehrens address conventional, paper-and-pencil tests used now in nearly all states to license teachers. Their chapter is aimed at helping those who commission, design, or use such tests to better understand their strengths and limitations and the inferences that might properly be drawn from them. The chapter discusses appropriate test development procedures, analyses of bias, the creation and interpretation of evidence concerning reliability and validity, and the reporting of test results. The chapter also illustrates that

while expert opinion diverges on some questions regarding test construction and use, a substantial consensus about appropriate testing practices exists.

Edward Haertel's chapter complements the preceding one by taking up the issues posed by "new" testing methods, such as performance tests and simulations for gauging teaching knowledge and skill. These methods are being explored by some states for use in licensure programs and by the National Board for future use in certification assessment. As Haertel notes, these methods "lie somewhere between observing teachers in their classrooms and tallying the circles darkened on their answer sheets." They may eventually be able to assess a greater range of abilities than traditional tests with greater reliability than classroom observations.

Haertel notes that one important rationale for investments in performance testing has to do with the effects that tests exert on the teacher-education curriculum. If tests more closely resemble the actual tasks of teaching, they might encourage improvements in teacher preparation while helping to codify the knowledge base for clinical practice. Haertel describes how performance exercises are used in assessment centers and professional examinations, and he offers examples of exercises being developed for teachers. He addresses the difficult issues of administering and scoring such exercises, representing and mediating conceptions of knowledge, and establishing reliability and validity.

If tests are well-designed, appropriately validated, scored, and administered, there remains the question of how to set standards as to what constitutes a passing score. Richard Jaeger's chapter is a practical guide to standard-setting that describes and discusses several methods used by states to set standards on tests used for admission to teacher-education programs or for certification. The chapter provides a comprehensive treatment of such issues as how to select and train the judges who will recommend or set standards, what the judges should do and based on what kinds of data. Jaeger addresses important matters of judgment along with relevant technical matters, recognizing the many kinds of real-world consequences that standard-setting entails.

Putting It All Together:

Combining Data for Decision Making

This section ends as it began with discussion of the need for multiple sources of evidence about teaching acquired by different methods. Chapter 19, by William Mehrens, addresses the issues associated with combining evaluation data from multiple sources, particularly when high stakes decisions — for example, hiring, tenure, licensure, or salary decisions — are being made. Mehrens describes several models for combining data, their assumptions, and their advantages and disadvantages for various purposes. He illustrates how the choice of a data-combination model will affect the decision reached. Readers will find here both a theoretical explanation and a concrete set of examples that illuminate a decision-making process more often conducted by instinct than analysis.

Part 3 — Cross-Cutting Perspectives

Any teacher-evaluation process exists within a broader organizational and social context. This final section of the Handbook treats some of the system-wide concerns that developers and users of evaluation plans must confront: legal and governance ramifications, ethical questions, economic trade-offs, and organizational factors that affect implementation.

In his discussion of legal issues, Michael Rebell points out that, ironically, even as the quality of teacher evaluation improves, more evaluation systems seem likely to be struck down by the courts. Whereas in the past, courts seemed more likely to defer to the judgments of school administrators, recent developments have led to increased court involvement. These include the larger number of teachers denied certification or continued employment because of increased emphasis on evaluation; the disproportionate effects of methods such as testing on minority candidates; the increased involvement of courts in social reform matters generally; and the increased experience and sophistication of judges in technical matters affecting litigation on employment discrimination and a variety of educational issues.

This chapter reviews state law affecting initial teacher certification, dismissal of probationary and tenured teachers, and performance evaluations; it also reviews the constitutional implications of these issues, generally litigated in federal courts. Rebell notes that the courts' normal propensity to favor objectivity and procedural regularity may push against the efforts of new assessment techniques that try to "keep faith with the integrity of teaching," but that an effective dialogue between courts, legislatures, educators, and researchers could enhance the prospects for professional standards that are both fair and faithful to the demands of teaching.

In chapter 21, Kenneth Strike takes up ethical issues concerned with teacher evaluation. He applies general ethical principles such as due process and equality to evaluation activities, producing a bill of rights that should govern the practice of teacher evaluation. Subsequently, Strike takes up questions of the sources of authority over the evaluation process. He describes three models of school organization — bureaucratic democracy, professionalism, and communitarian democracy — and uses them to explore issues of governance and political values as they affect and are manifested in school activities such as evaluation. These models give different weights to the roles of bureaucratic authority, professional knowledge, and parental and community interests in the operations of schools. Strike notes that any view of educational governance and of the institutional forms that will structure evaluation practice must respect the central political values of each model and must integrate and balance them in responsible ways.

These governance issues are addressed even more explicitly by Arthur Wise and Tamar Gendler, who examine the interface among public interests, employer concerns, and professional responsibilities across the broad spectrum of teacher

evaluation occasions: licensure, selection, tenure, continuing professional development, career advancement, and school improvement. Wise and Gendler sort through how different governmental units and professional bodies are and ought to be involved in each of these arenas based on an assessment of how public and client interests may best be served. They distinguish among professional bodies (such as professional standards boards and their analogs) at the national, state, and local levels, and between state and local governments — legislatures and executive agencies, school boards and school personnel — in analyzing the legal, political, and professional imperatives for allocating authority and responsibility. Drawing on legal and organizational principles, and the experiences of other professions, this chapter illuminates the rationales for allocating roles and responsibilities for various evaluation functions to the parties most able to discharge them effectively.

Stephen Hoenack and David Monk treat the economic aspects of teacher evaluation. They call our attention to the ways in which evaluation systems can themselves be evaluated with respect to how efficiently they use resources and how effectively they create useful incentives. Among the concepts they bring to bear on the questions of evaluation are production theory — how evaluation may structure educational inputs or outputs by the criteria it emphasizes and the factors it assesses; the economics of information — how to assess the costs and benefits of information that might be gathered during an evaluation process based on its contribution to evaluation goals; and performance incentives — how evaluation systems produce both intended and unintended incentives and how planned incentives may sometimes serve as alternatives to traditional evaluation practices.

The authors also discuss the distributional effects of policy interventions. Among other things, they note that evaluative information can have differential impact on the welfare of various groups involved in the educational process; for example, taxpayers, teachers, administrators, and students. Equally important, a teacher evaluation system could work to the advantage of one type of student and to the disadvantage of another, particularly if it creates incentives to change the relative amounts of teacher effort devoted to different categories of students (or to categories of activities that differently affect students). Recognition of these issues should help policy makers, school administrators and teachers think ahead to the likely consequences of decisions that structure evaluation criteria and processes.

Milbrey McLaughlin applies this range of cross-cutting perspectives to the implementation process itself in her chapter, "Embracing Contraries: Implementing and Sustaining Teacher Evaluation." In a comprehensive look at the real-world problems of evaluation, McLaughlin notes that technical and management concerns converge where organizations must create a culture that can accept and support evaluative activity. Such contraries as accountability and improvement, bureaucratic control and professional autonomy, individual needs and institutional goals are presented throughout this volume and confronted here.

While these competing needs bedevil proponents of teacher evaluation and may sometimes defeat implementation efforts, they are essential to a productive educational system. McLaughlin describes how organizational commitment and resources, joint training for administrators and teachers, integration of individual and institutional improvement activities, and strategies for ongoing review and revision of the evaluation process can contribute to a culture that supports and sustains evaluation, embracing contraries with common commitments among teachers and administrators.

The Handbook's final chapter reflects and summarizes many of the principles and perspectives presented in the preceding chapters. Daniel Stufflebeam and James Sanders provide an overview of the *Personnel Evaluation Standards* developed by the Joint Committee on Standards for Educational Evaluation, representing 14 major organizations of professional educators, researchers, and government officials. The *Standards* are designed to assist school districts, universities, and other agencies in examining and improving their systems for evaluating professional personnel.

The chapter discusses the four basic attributes judged by the Committee as essential to all evaluations. These include (1) *proprietary standards*, aimed at ensuring that the rights of all persons affected by an evaluation system are protected; (2) *utility standards*, intended to guide evaluations so that they will be informative, timely, and useful; (3) *feasibility standards*, intended to increase the prospects for practical, politically viable, and fiscally viable evaluation systems; and (4) *accuracy standards*, designed to ensure that the information obtained will be technically defensible and the conclusions drawn will be supported by that data. Stufflebeam and Sanders also describe how the standards may be applied to teacher evaluation systems in balancing the complex, sometimes competing, concerns that inevitably arise within organizations striving to orchestrate the energies of teachers, administrators, parents, and students on behalf of common goals.

PART I

Purposes of Evaluation

2. Intended and Unintended Consequences: Purposes and Effects of Teacher Evaluation

Gary Natriello

The aim of this chapter is to provide a broad framework in which to consider the various purposes and effects of teacher evaluation in school organizations. The distinction between purposes and effects employed here is a quite straightforward one. *Purposes* are those reasons for which the evaluation process is initiated. *Effects* are those results of the evaluation process. Effects may or may not be related to the initial purposes of the process, but they are always related to the activities or practices undertaken as part of the evaluation process.

Recognition that teacher evaluation occurs within an organizational context is central to understanding the major purposes of teacher evaluation systems. The evaluation of the performance of individual teachers is performed constantly, by students, by parents, by other teachers, and by the general public. But only when the teacher-evaluation process is tied to the school organization do we deem it a system and attempt to study its formal properties and examine the intended purposes and consequent effects. Thus, it is only these formal systems of teacher evaluation that are the topic for this chapter.

Teacher-evaluation systems are only one of several kinds of control processes in schools. Control is also exercised through such things as rules and procedures, available technology, and professional norms. But teacher evaluation serves the purpose of organizational control through several different functions with several different levels of effects. It is an overall understanding of these multiple purposes and effects that this chapter attempts to develop. An understanding of this sort will enhance our capacity to assess the total impact of teacher-evaluation processes and explain the design and operation of what at times may seem to be arbitrary or irrational practices.

AUTHOR'S NOTE: Work on this paper was supported by grants from the Spencer Foundation. The author would like to thank Sanford Dornbusch, Daniel Duke, and Michael Knapp for their helpful comments on an earlier draft of the paper.

Three Purposes of Teacher Evaluation in Schools

There are any number of ways to classify the purposes of teacher-evaluation processes, but teacher evaluation serves at least three major purposes in contemporary schools. The differences among these three purposes are most apparent when their impact is considered at the level of the individual teacher, particularly when an attempt is made to bring about some change in the status quo.

First, evaluation may be used to control or influence the performance of individuals within particular positions. Such evaluation processes are designed to maintain current performance or bring about changes in performance without any changes in the status of the performer (Kearney, 1978). Any change in individual performance would be the result of performing the same set of tasks at a different performance level.

Much of the evaluation activity that occurs within a school involves attempts by principals and other evaluators to influence the performance of individual teachers in their current assignments (e.g., Lortie, 1969; Embretson, Ferber, & Langager, 1984). The goal is to improve performance that is already within a range that is considered acceptable for holders of the position. There are a great many quite different approaches and techniques for accomplishing this end. For example, teacher-evaluation systems that use criteria based on effective teaching research (Beach & Reinhartz, 1984), those employing models of clinical supervision (Goldsberry, 1984), and those utilized to determine merit pay increments (Natriello & Cohn, 1983) may all be classified as attempting to improve the performance of individual teachers in their current assignments.

Second, evaluation may be used to control movement into and out of positions. Evaluation may serve to screen individuals attempting to enter a position, to retain individuals in a position, or to enforce the exit of individuals from a position. Such evaluation processes are designed to change an individual's performance by changing the tasks performed or to change the performance of a system by changing the individuals assigned to perform certain tasks (Tenopyr & Oeltjen, 1982).

There have been many discussions and many policy reforms regarding the use of evaluation to make decisions about the status of teachers. Many states have recently revised processes for determining eligibility for certification and many of these changes have involved the institution of new forms of evaluation as requirements for certification. For example, a growing number of states now use some type of examination for individuals seeking to enter teaching (Pugach & Raths, 1983; Anrig, 1986).

The actual hiring process is another stage at which evaluation may be used to make decisions about admission of an individual into the position of teacher. Individual districts employ various assessment processes as mechanisms to screen candidates for teaching positions. Once individuals are in teaching positions in districts additional personnel decisions may be based on evaluations. The granting or denial of tenure is typically based on some evaluation process,

often the same process used to help teachers improve their performance. Decisions about the movement of teachers on career ladders are also typically tied to some evaluation process. Finally, the dismissal of teachers, tenured or not, whose performance is deemed unacceptable might be based on evaluations of performance (Bridges, 1984).

In all of the cases the key element is that evaluation is used to change or maintain the performance of the system, not by changing the performance of individuals who remain in their current positions, but by changing the individuals who are in the positions. In contrast to the first purpose of improvement, in this case evaluation leads to movement.

Third, evaluation may be used to legitimate the organizational control system itself. Evaluation may serve to convey a sense of justice and equity both about the organization and about its control of attempts. Such evaluation processes are designed to influence performers by convincing them that the dictates of the evaluation process, for either a change in the performance of an individual in a position or a change in the position of an individual are legitimate and deserving of compliance.

Evaluation systems are key organizational processes for mediating the conflict between the hierarchical authority system of the modern organization and the democratic traditions of U.S. society. In order to be acceptable to subordinates, evaluation systems must exhibit certain features consistent with democracy even though they function as systems for one group of individuals to control the behavior of another group. For example, in U.S. society there is an expectation that all individuals will be treated equally. Evaluation systems must offer equal or fair treatment if they are to be perceived as legitimate (Peterson & Peterson, 1984). Evaluation systems must also conform to the requirements of due process (Strike & Millman, 1983). Evaluation systems thus meet the expectations of organizational members and the society at large that organizations will exhibit rational decision-making processes (Meyer & Rowan, 1977).

Thus, teacher evaluation processes may be intentionally used by school managers: (1) to influence the performance of individuals who will remain in their current positions; (2) to guide decisions about the movement of individuals into and out of positions; and (3) to legitimate the control attempts of the school organization. While these functions have associated with them certain anticipated effects, there are a variety of other effects of teacher evaluation processes that are neither intended nor always apparent to those planning and operating evaluation systems.

Three Levels of the Effects of Teacher Evaluation

Because teacher evaluation has a broad range of effects, both intended and unintended, it is useful to provide some method of classifying these effects. One way to classify them is in terms of the level at which the impact falls. Teacher

evaluation has effects on the individual, organizational, and environmental levels. Individual-level effects occur when evaluation processes or practices employed in the evaluation of a particular teacher have some impact on that teacher. Organizational-level effects occur when the evaluation processes or practices in a school have an effect on teachers in the school. Such effects may stem from evaluation processes employed in the school in general or those employed in the evaluation of a particular colleague. Moreover, such effects may pertain to the reactions of one teacher, or all teachers other than the one being evaluated directly, or to a general feature of the organization such as consensus about goals or level of communication activities. Finally, effects at the environmental level occur when the processes and practices of teacher evaluation, either as they pertain to a single teacher or to the school as a whole, have some impact on individuals and institutions outside the immediate school organization. Thus, although every evaluative act has a direct target or audience (i.e., the particular teacher or teachers being evaluated), each evaluation activity also has an audience of others within the school and an audience of those outside the school. For example, the dismissal of a teacher for reasons of incompetence and the delineation of the incompetent performance has an effect not only on the teacher in question but also on others within the school and on the larger community.

Individual Level Effects

Effects at the individual level are the most direct and the most likely to be anticipated. Individual performers subject to the evaluation process are affected by it. For example, individual teachers may react to the evaluations of their own performances in various ways, only some of which might have been intended by the designers and managers of the evaluation system.

One purpose of evaluation systems is to lead to the improvement of the performance of individual teachers or the maintenance of already acceptable levels of performance. Presumably, teacher performance can be enhanced if evaluators provide feedback on problems in performance and strategies for remedying these problems. Thus, the communication of information about performance may have positive effects, particularly for teachers who most often practice out of sight of other adults and so rarely have the benefit of the perspective of another professional educator. The impact of information on teaching performance might operate in one of two ways. On the one hand, it might be assumed that simply pointing out problems in teaching to a teacher would lead that teacher to improve. This seems to be an assumption of evaluation systems that provide feedback on performance without much attention to organized follow-up activities to facilitate improvement. On the other hand, it might be assumed that information on problems with a teacher's performance will produce little improvement by themselves in the absence of professional development opportunities. This seems to be the assumption of the small number of schools and districts where evaluation activities are closely connected with staff

development opportunities. No doubt there are some aspects of teaching that fit each assumption, although most complex improvements in practice are likely to require organized development opportunities. In either case the primary effect of the evaluation process is to provide information on performance to foster development and improvement.

Improvement, of course, is in the eye of the beholder. Teachers may change their teaching so that it is more pleasing to their supervisors and/or more effective in their own eyes. The former may be accomplished by bringing their performance of aspects of the teacher's role or the balance of effort among various aspects of that role into line with school and district expectations. The latter may be achieved by altering their teaching to improve the results they achieve for a given level of effort. In reporting on a study of the evaluation of teachers in urban middle schools, Natriello (1983) argues that when teachers achieve greater results for a given level of effort, they enhance their leverage over the tasks they confront and their control over their work. This study suggests that increased evaluation activity can lead to this enhanced self-perception of leverage.

But the teacher-evaluation process may function in other ways that lead to the improvement of performance. Evaluation processes may have motivational as well as instrumental consequences. Teaching is an occupation in which individual teachers typically perform alone and out of sight of other adult colleagues. Moreover, the nature of teaching leaves teachers feeling quite uncertain about their teaching performance and its effects on students (Lortie, 1975). Evaluation activities can provide teachers with both contact with other professionals in the school and with reassurance about their performance. Teachers who are seldom evaluated feel isolated and undervalued (Natriello & Rowe, 1981; Natriello & Dornbusch, 1981). Teachers in some schools have so little contact with other professionals that Natriello and Dornbusch (1981) found that teachers who were infrequently evaluated actually preferred more frequent evaluations even when those evaluations produced negative outcomes.

Teachers who are well regarded also appear to desire more frequent evaluation. In a case study of an elementary school, Natriello and Rowe (1981) reported that teachers who were well regarded by the principal complained that they still needed to know the principal's evaluation of their teaching more frequently to feel good about what they were doing. The impact of evaluation on motivation is captured by the report of a teacher in this same elementary school. This veteran teacher, who was also held in high esteem by the principal, explained to the researchers that he had recently been honored by the parents and the staff with a celebration to mark his impending early retirement. With tears in his eyes, he reported that if he had any notion that people thought what he was doing was valuable he would have continued teaching.

Teachers may interpret the evaluations they receive as a challenge to perform at a higher level. But the motivational effects of teacher evaluation cut both ways. While some teachers may be challenged by evaluations, others may

experience extreme stress and anxiety which is never converted to improved performance. Individual differences in teacher responses to evaluation are difficult to predict and address, but differences that are related to specific evaluation practices are both more predictable and more easily addressed by evaluation policies. Teachers appear less likely to react negatively to evaluations when they have a prior understanding of what is expected of them, when adequate information is collected on their performance, when feedback is frequent and informative, and when resources are provided to help them improve (Natriello, 1983). Teachers react negatively to an evaluation process that identifies problems that cannot be corrected, given the working conditions they must confront (Natriello & Cohn, 1983).

Teacher evaluation in order to make decisions about the status of teachers has an obvious effect on the individual level when it results in the entrance, maintenance, or removal of an individual in a given position. Entry or removal from a position require teachers to assume or renounce a role and the tasks and responsibilities that come with it.

In addition to the effect of the actual change of position or status, there are also effects from the anticipation or prospect of a change of status, characterized by either challenge or fear. Individuals are typically aware of the linkage between evaluation and status decisions and so may strive to perform well on activities subject to evaluation to maximize their chances of attaining or retaining a desired status. Such effects may operate even if the individual in question never experiences a change in status. There are several common examples of such effects. Similarly, for those anticipating entry into a position, the prospect of the new position and its responsibilities may lead to changes in current performance and thinking to bring them into line with the anticipated position.

Practices associated with the use of teacher evaluation to legitimate the control system of the organization also have direct effects on individual teachers. Evaluation systems vary in the extent to which teachers perceive evaluations of their performance to be soundly based, that is, reflective of their true performance. The more teachers perceive evaluations of their performance to be sound, the more legitimate they will deem the evaluation system and the more effort they will devote to tasks upon which they are evaluated (Dornbusch & Scott, 1975; Natriello & Dornbusch, 1981, 1984). Evaluation practices that tend to produce sound evaluations of teaching performance such as collecting more representative samples of teacher performance are thus likely to lead to greater teacher effort.

Evaluation systems also vary in the extent to which teachers perceive evaluations to be consistent from evaluator to evaluator. The more teachers perceive the evaluation system to be consistent, the more likely they are to perceive it as just, and the more effort they will devote in response to the evaluations they receive (Natriello & Dornbusch, 1984). Evaluation practices that tend to produce more consistent evaluations from evaluator to evaluator, such as more well-defined

guidelines for linking performance to assessment, are thus likely to lead to greater teacher effort.

Organizational Level Effects

As noted earlier, *organizational level effects* are those that involve others in the school aside from the individual teacher being evaluated directly. These effects may result from the evaluation of one teacher or from the evaluation system in general, and their impact may involve one teacher or the entire school. Members of the school organization are inevitably observers of the evaluation processes that occur within the organization. For example, teachers in a school may perceive and react to the evaluations received by their colleagues. Such evaluations within the school organization may communicate the norms for performance in the school even to those whose performance is not being evaluated directly. When one teacher sees another being negatively evaluated for a certain kind of performance, the first teacher learns to avoid that kind of performance. Similarly when certain teachers are rewarded with good evaluations for their performance, other teachers learn that such performance is expected. In this way the evaluation system has effects that lead to the improvement of performance in the school overall, independent of direct evaluations of teacher performance. Of course, if the evaluation process is not carefully executed and evaluations are not soundly based, then others teachers observing this process will discount any connection between their performance and the evaluation they might receive. In any case, these effects are organizational-level effects because they have an impact on a whole class of organizational performers.

The effects of evaluation processes on other teachers may vary with the norms governing the sharing of information about one's evaluation. For example, Natriello and Cohn (1983) report on a district in which teachers shared information on merit increments at the secondary level, but kept such information confidential at the elementary level. In this case, the impact of the evaluation process on other teachers might be expected to be greater at the secondary level.

Evaluation activities for the purpose of fostering improved teacher performance also have effects at the organizational level that are more general. For example, Natriello and Wilson (1986, 1987) investigated the relationship between evaluation practices and such organizational dimensions as goal consensus, communication, and leadership supportive of professional growth for teachers. They found that in schools where evaluation activities were more evident, these other dimensions were also more evident, though the direction of the effect could not be determined. Nevertheless, it is not unreasonable to expect that teacher-evaluation practices might lead to greater goal consensus among the staff of a school. Such evaluation practices might also lead to the diagnoses of organizational problems if information collected for the evaluation of individual teachers is aggregated and examined as evidence of school-wide phenomena.

There are other organizational-level effects of teacher-evaluation systems. A teacher-evaluation system developed with a high level of teacher participation may lead to shared discussions of educational issues and greater communication among the members of a school staff about their teaching. For example, Natriello & Cohn (1983), in their case study of a district with an evaluation system linked to a merit pay plan, noted that the formal participation of teachers and their representatives in defining the criteria for good teaching led to a good deal of discussion among teachers about the most appropriate and effective teaching practices. In addition, more active teacher-evaluation systems may lead teachers to adopt more consistent approaches to educational processes in their schools as they all come to be influenced by the criteria and standards of the evaluation system.

Evaluations connected with entry to, stability in, and exit from teaching positions also have effects throughout the school organization. As teachers observe changes in the status of colleagues that stem from evaluation processes, they come to a clearer understanding of their identity as members of the school, that is, they come to know the qualities that separate members from nonmembers. By doing this they come to understand more clearly the values and norms of the school; they learn their roles in the school organization and the school's role in the community. Such boundary-strengthening processes come to define more sharply the expectations for teachers and teaching performance.

There are also organizational-level effects related to decisions about entry to, stability in, and exit from teaching positions that operate through the composition of the teaching staff in a school. Changes in the composition of the staff, that is, in one or more members of the group of teachers, can lead to a wide array of organizational-level changes. For example, a new teacher with special curricular expertise may join the group and permit the school to offer a new dimension in curriculum. Or a new teacher able to devote greater effort to the school and to teaching may set in motion new competitive forces that lead other teachers to increase their effort in teaching and other school assignments. Or a new teacher may collaborate with current teachers to enrich the teaching of all concerned. Finally, when a decidedly poor teacher is dismissed from the staff, control mechanisms required to keep the poor performer in check may be eliminated, thereby increasing the autonomy of the remaining teachers.

In addition to the organizational-level effects of evaluation conducted for the purposes of improvement and movement, there are also such effects from evaluation conducted for the purpose of enhancing the legitimacy of the control system. The evaluation system may also communicate norms regarding fairness and membership to those in the school organization. If teachers see their colleagues being evaluated according to a consistent set of criteria and standards, they may come to believe that the organization will treat all teachers in an equitable manner. On the other hand, when teachers believe that some principals give lower evaluations than others, then they come to believe that their evaluations depend more on the evaluator than on their performance as teachers. A

rational response to such working conditions would be to devote efforts to seeking a transfer to another principal rather than to improving one's teaching performance (Natriello & Rowe, 1981).

Similarly, if teachers perceive that decisions regarding the status of their colleagues are made on the bases of rational evaluation processes, they may feel more secure in that their own positions are safe from arbitrary administrative decisions. This may add to the legitimacy of the evaluation system and the organization in the eyes of the teachers. Such conditions may enhance the commitment of teachers to the school. When teachers feel that decisions regarding their future may be based on arbitrary administrative decisions over which they have no control, they are more likely to seek other employment opportunities.

Environmental Level Effects

A third type of effect of evaluation processes may be called *environmental level effects*. Such effects are those that occur in the environments of the schools. Those outside the immediate school organization are yet another audience for the operation of evaluation systems. For example, a school district with well-articulated policies for the evaluation of teachers may be viewed by those outside the organization as more professional or modern.

Many of the effects of teacher-evaluation systems on those outside the immediate school organization involve other organizations. For example, state and national teachers organizations, state education agencies, governmental bodies that regulate the treatment of employees, and the courts (Strike & Bull, 1981) all have reacted to teacher-evaluation systems from time to time. School districts are often in a position of having to explain and justify their approach to the evaluation of teachers to these outside agencies that have a specific interest in education and teachers. At times these organizations make comparisons among the teacher-evaluation systems operated by various districts and render judgments, formal and informal, as to the adequacy of such systems.

Others outside the school with a more general interest in education may also react to teacher evaluation practices. For instance, members of the business community may question the personnel policies of a school or district if the teacher-evaluation system is missing certain key elements. Parents and other members of the local community may also question the evaluation practices of schools, particularly when they are related to specific personnel decisions that may be controversial. Teacher-evaluation systems may either contribute to or alleviate such controversies.

There are dramatic differences in the visibility of the three functions of teacher-evaluation systems at the environmental level. Most notably, the specifics of the improvement function of teacher evaluation are deliberately not made visible to certain individuals and organizations outside the school. Parents are denied information on the improvement related activities of teacher-evaluation

systems because to do otherwise would be to acknowledge that a particular teacher, perhaps the teacher of their children, is in need of improvement. Even boards of education receive little information on the needs of staff members for improvement.

The emphasis instead is on the status decisions that might stem from evaluation processes. Hiring, retention, and dismissal are highly public acts justified by evaluation activity. While improvement activities are typically kept between teachers and other immediate professional staff, status decisions are often reviewed by more people in the district and outside the district.

The review of such status decisions is often made to determine the legitimacy of the actions of the organization. Thus, environmental-level effects generally concern the status determining function and the legitimation function of teacher evaluation.

Conclusions

While it may not be possible to anticipate all of the effects of any particular set of teacher-evaluation practices, consideration of the impact of teacher evaluation at the individual, organizational, and environmental levels does permit examination of a wide range of possible effects. As the examples above have illustrated, the effects at each level may be positive or negative, depending upon the particular practices employed for the evaluation of teachers. Moreover, the same practices may produce positive effects at one level and negative effects at another level.

The variety of effects of teacher evaluation discussed above obviously do not all occur all the time in all districts that engage in teacher-evaluation activities. Indeed, the particular set of effects in a school or district would depend upon both the specific teacher evaluation practices employed and the local context. In some schools where the purposes of teacher evaluation are more limited than the three described above, certain effects would not be felt. In other schools, particular evaluation practices may lead to one of two quite different outcomes. For example, certain districts may operate evaluation systems in such a way as to enhance the legitimacy of the control process, while in others, the evaluation practices might diminish the legitimacy of the control system.

An awareness of the potential effects of teacher-evaluation processes will not guarantee successful outcomes, but it will sensitize those involved in teacher-evaluation systems to the processes and effects that their activities may set in motion. Because the unintended consequences of teacher-evaluation systems are just as real in their impact as those that are intended, sensitivity to the full range of outcomes is essential for a complete understanding of the impact of teacher evaluation.

References

- Anrig, G.R. (1986). Teacher education and teacher testing: The rush to mandate. *Phi Delta Kappan*, 67, 447-451.
- Beach, D.M. & Reinhartz, J. (1984). Using criteria of effective teaching to judge teacher performance. *NASSP Bulletin*, 68, 31-37.
- Bridges, E.M. (1984). The identification, remediation, and dismissal of incompetent teachers. *School Management Digest*, 2, 1-57.
- Dornbusch, F.M. & Scott, W.R. (1975). *Evaluation and the exercise of authority*. San Francisco: Jossey-Bass.
- Embretson, G., Ferber, E., & Langager, T. (1984). Supervision and evaluation: helping teachers reach their maximum potential. *NASSP Bulletin*, 68, 26-30.
- Goldsberry, L.F. (1984). The realities of clinical supervision. *Educational Leadership*, 41, 12-15.
- Kearney, W.J. (1978). Improving work performance through appraisal. *Human Resource Management*, 17, 15-23.
- Lortie, D. (1969). The balance of control and autonomy in elementary school teaching. In A. Etzioni (Ed.) *The semi-professions and their organizations*, (pp. 1-53). New York: Free Press.
- Lortie, D. (1975). *Schoolteacher: A sociological study*. Chicago: University of Chicago Press.
- Meyer, J.W. & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83, 340-363.
- Natriello, G. (1983). *Evaluation frequency, teacher influence, and the internalization of evaluation processes: A review of six studies using the theory of evaluation and authority*. Eugene, OR: Center for Educational Policy and Management, University of Oregon.
- Natriello, G. & Cohn, M. (1983). *Beyond sanctions: The evolution of a merit pay system*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.
- Natriello, G. & Dornbusch, S.M. (1981). Pitfalls in the evaluation of teachers by principals. *Administrator's Notebook*, 29, 1-4.
- Natriello, G. & Dornbusch, S.M. (1984). *Teacher evaluative standards and student effort*. New York: Longman.
- Natriello, G. & Rowe, B. (1981). *Life in a loosely-coupled world*. Unpublished paper, Washington University, St. Louis, MO.
- Natriello, G. & Wilson, B. (1986, April). *Teacher evaluation as an organizational process*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Natriello, G. & Wilson, B. (1987, April). *Teacher evaluation and school climate*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Peterson, D. & Peterson, K. (1984). A research-based approach to teacher evaluation. *NASSP Bulletin*, 68, 39-46.
- Pugach, M.C. & Raths, J.D. (1983). Testing teachers: Analysis and recommendations. *Journal of Teacher Education*, 34, 37-43.
- Strike, K. & Bull, B. (1981). Fairness and the legal context of teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation*. Beverly Hills, CA: Sage.
- Strike, K. & Millman, J. (1983). Nontechnical questions about teacher evaluation systems in elementary and secondary schools: A research agenda. *Educational Evaluation and Policy Analysis*, 5, 389-397.
- Tenopir, M.L. & Oeltjen, P.D. (1982). Personnel selection and classification. *Annual Review of Psychology*, 33, 581-618.

3. Preservice Evaluation of Teachers

Walt Haney

I. Introduction

How are prospective teachers evaluated prior to their actual entrance into the occupation of teaching? This question is the focus of this chapter. Specifically treated is the topic of preservice evaluation of teachers in terms of admission of individuals into teacher-training institutions and programs, and requirements for continuation and graduation from teacher-training programs. The topic of evaluation via special teacher exams (so-called teacher-competency tests) is not extensively treated, as it is dealt with elsewhere in this volume (see also Haney, Madaus & Kreitzer, 1987). Before embarking on this agenda, several prefatory remarks are offered to clarify the nature and importance of preservice evaluation and screening of prospective teachers.

II. The Nature and Importance of Preservice Evaluation of Teachers

Perhaps the most universal requirement for entering into the occupation of teaching in the United States, at least since midcentury, is the possession of a college degree. For this reason, requirements connected with the earning of a college degree — things such as getting admitted into college, earning passing marks in college courses, and meeting graduation requirements — are closely connected with the preservice evaluation of prospective teachers. These matters will be treated later in this chapter, but by way of introduction, it should be noted that requiring school teachers to have graduated from a four-year college is a relatively recent development in the social history of the United States. As late as 1917, roughly half of the nation's elementary and secondary teaching force had "no more than four years' education beyond the eighth grade" and "had no

AUTHOR'S NOTE: The author would like to thank Mary Kennedy and Marianne Castano for help with and suggestions for the preparation of this chapter.

special professional preparation for the work of teaching" (Evenden, 1933, p. 32). It is only in the last 50 years that college education has become inextricably intertwined in the selection and evaluation of teachers in the United States.

A second general point worth noting is that in twentieth century America, there have been repeated crises over teacher shortages. History clearly shows that market and social forces severely condition the impact that any formal mechanisms of teacher selection and evaluation have on the recruitment and retention of teachers. As public schooling rapidly expanded in the first two decades of this century, during and just after the first world war, for instance, the nation experienced a severe teacher shortage. During the Depression there was a surfeit of teacher candidates, but in the manpower shortage brought on by World War II, there was again a shortage of teacher candidates. As Sedlak and Schlossman (1987) comment, "Contrary to what many modern-day educators tend to assume, teacher shortages have been commonplace throughout the twentieth century" (p. 123). Another student of the supply and demand for teachers, Weaver (1983, 1984), argues that given the dynamics of the teacher labor market and the political economy of teacher training institutions, various proposals to upgrade standards for teachers, for example via testing of prospective teachers or by changing certification requirements, hold little promise of upgrading teacher quality generally, unless the incentives for people to become teachers are improved and the missions of schools of education are redefined.

What this suggests, of course, is that formal mechanisms for evaluating prospective teachers have little potential for affecting the composition of the teaching work force when there are fewer teacher candidates than there are openings for new teachers. For it must be remembered that the demand for teachers is highly inelastic relative to the labor market supply of prospective teachers. States that require students to attend school have a concomitant obligation to supply people to staff the classrooms that children are legally required to attend. Thus, it is not surprising that during periods of teacher shortages, many have been hired as teachers who have been only temporarily or provisionally certified. Hence, it is important to note at the outset of this review that during periods of teacher shortage, broad social and labor market forces may influence the supply of teachers to a much greater extent than any formal mechanisms for evaluating teachers — at either the preservice or inservice levels.

Nevertheless, preservice screening of prospective teachers is still highly important. Only 70%-80% of recent 17- to 19-year-olds have been graduating from high school with their age group cohorts (Plisko, 1983, p. 58-59; Stern, 1988, Vol. 1, p. 29). And college freshmen enrollments amount to only about 80% of high school graduates (Plisko, 1983, p. 90). Moreover, bachelor's degree graduates in recent years constitute roughly 30%-40% of those who enroll as freshmen (Plisko, 1983, p. 116, *Digest of Education Statistics*, 1988, pp. 141, 175). This means that for every 100 high school students who potentially could become teachers, 75 to 85 will not progress far enough through the education pipeline to achieve a bachelor's degree from a four-year college — a credential

that has become a virtual minimum requirement for certification as a teacher, either temporary or permanent, anywhere in the United States. This suggests that despite a great deal of attention focused recently on certification tests for people about to enter teaching (or in a few states for those who want to be recertified), from a broader labor market point of view, the processes by which people are screened and evaluated long before even becoming eligible to sit for such tests are far more influential in determining who will teach than are the last few steps in teacher-selection-and-evaluation processes.

Hence, it is now time to examine some of the broad features of preservice evaluation and screening of people who might become teachers; namely standards for admission into teacher-training programs, grading and course requirements of students in them, and standards by which students become eligible to graduate from them. As practice teaching is a requirement for almost all preservice teachers and is a requirement quite different from those for most bachelor's degree programs, a separate brief section is devoted to evaluation of practice teaching.

III. Admission into Teacher Education Institutions

As previously mentioned, early in the twentieth century a majority of public school teachers in the United States had less than four years of education beyond the eighth grade. Up until then, local and county exams had been the prime means by which teacher candidates were selected and evaluated. However, during the first two decades of this century, there was widespread criticism of the practice of evaluating and certifying teachers on the basis of local and county exams, and formal teacher training was touted as an alternative to certification via examination. During the first third of the twentieth century there was "an increased tendency to substitute preservice preparation for state and county examinations as bases for certification of teachers, and a marked increase in the number of normal schools which extended their curricula from two to four years" (Evenden, 1933, pp. 30-31).

Gradually, higher education in special-purpose, teacher-training institutions came to be replaced by preparation of teachers in institutions of higher education more generally. During the 1950s there remained over 100 specialized teachers colleges. But with widespread demand for higher education in the 1950s and 1960s, most teachers colleges were turned into more general-purpose state colleges. In 1967 only 9 state teachers colleges remained as specialized institutions of higher education. As Stinnett (1967) commented, "Clearly, teacher education has become an integral part of the mainstream of higher education," with approximately 1,200 college and universities nationwide engaged in teacher preparation.

As teacher education became part of the mainstream of higher education, entry into teacher-education programs were probably affected more by colleges'

and universities' general admission requirements than by teacher-education admissions requirements per se. College admissions requirements refer, of course, to such things as high school graduation, college admissions tests, letters of reference, personal statements, and in more selective institutions, personal interviews.

It is difficult to gauge the effects of such general college admissions requirements on the pool of young people entering teacher education programs. It is clear that degree of selectivity of colleges and universities generally has varied over time and demographics. For example, while demographics led to a decrease in the number of students graduating from high school from the mid-1970s to the mid-1980s (from 3,155,000 in 1975-76 to 2,890,000 in 1982-83, or a more than 8% decline, *Information Please Almanac*, 1987, p. 824), the number of institutions of higher education nationwide (and the number of faculty in them) continued to increase (Grant & Snyder, 1983, p. 101). Thus, it was surely easier for a high school graduate to win admission into some kind of institution of higher education in the 1980s than it was in the 1960s and early 1970s. As Willingham and Breland (1977) commented, "The existence of ample (perhaps too ample) staff and physical resources, steady or declining applicant population and serious financial problems in many institutions has tended to create a buyer's market in undergraduate admissions" (p. 75).

Amidst the more than 3,300 institutions of higher education in the United States in the 1980s, (Snyder, 1987, p. 11), admissions standards vary widely. Some undergraduate institutions practice selective admissions, but many others admit virtually all who apply (Willingham & Breland, 1977). Astin, Green, Kern, and Schalit, for example, in developing a sampling strategy for the 1986 survey of college freshmen conducted by the Cooperative Institutional Research Institute, estimated that there were 25 private universities with average combined Scholastic Aptitude Test Verbal and Math scores (SAT V + M) of more than 1,175, but there were also 123 four-year Protestant colleges for which SAT V + M average scores for incoming freshmen were less than 875 (or were unknown, p. 108). So too do admissions standards vary widely among the more than 1,200 colleges and universities that engage in teacher preparation. Moreover, it is clear that market forces mentioned above have had a dramatic impact on the pool of young people who might seek admissions into teacher-preparation programs. According to annual surveys of incoming college freshmen over the last 20 years, the proportion of freshmen expressing a preference for a career in elementary/secondary teaching has fallen precipitously from nearly 25% in 1968 to a low of only 4.7% in 1982 — though the proportion expressing a career preference for teaching has edged upwards to 7.3% in 1986 (Astin et al., 1986, p. 6). Clearly, whatever the intended standards for admissions into a teacher-education program, they would almost surely be applied differently to an applicant pool of 250 than to an applicant pool of only 42 or 73.

Nevertheless, there has been a long tradition of criticism of teacher-training programs as having lower admissions standards than other higher education

programs. In the Pennsylvania Study, carried out from 1928 to 1932 by William Learned and Ben Wood, for example, it was found that high school seniors going into teachers colleges scored more poorly on tests of science and math than high school students they might well have to teach once they finished their teacher-training programs (Resnick, 1982; Haney, 1984). Recently, related surely to the decrease in numbers of students interested in elementary/secondary teaching, there has been an upsurge in criticisms of admissions standards for teacher-training programs. In 1979, Weaver presented data from ACT and SAT college admissions tests to support his contention that "schools of education are selecting potential educators from among the least academically talented populations applying for college admissions" (p. 29). In 1982, Hodges and Nash said that "recent research has shown that the majority of students seeking admission to certain teacher education programs were writing at or below the ninth grade level . . . while education majors' reading competency scores declined from 1967 to 1977, especially among prospective elementary school teachers" (Hodges & Nash, 1982, abstract). In 1984, Feistritzer's report on teacher education recommended that up to half of the 1,287 teacher-education programs in the United States should be closed because of low or nonexistent entry and exit standards. And in 1986, Tanner opened his paper, "Teacher Education and Academic Rigor," by observing "There is a common assumption of nonrigor associated with teacher preparation programs. They are perceived as dumping grounds for college students who lack either the ability or the initiative to compete effectively in other traditional academic programs" (p. 1).

Such criticism of standards for admission into teacher-education programs has led to two sorts of responses. First, a minor spate of publications have sought to refute the idea that people admitted into teacher-education programs are somehow less academically talented than those admitted into other higher education programs (e.g., Ishler, 1984; Nelli, 1984; Tanner, 1986).

Second, a variety of proposals has been advanced to upgrade requirements for admission into teacher education programs, independent of criteria for admissions into higher education more generally. Perhaps most prominently, in 1985, the National Counsel for Accreditation of Teacher Education (NCATE) issued new "standards for accreditation of professional education units" (i.e., colleges, school departments, or other administrative bodies responsible for "the preparation of students who seek state certification as teachers or other professional education personnel." (NCATE, 1985, p. 1). Among *preconditions* for institutions to establish eligibility for NCATE accreditation evaluation are that:

The unit assesses the basic skills of education students using standardized tests and other appropriate measures prior to admission to the program (NCATE, 1985, p. 6).

Also, however, a number of states have mandated new standards for admissions into teacher-education programs, with standards defined in terms of general college admissions tests, grade point averages and scores on special teacher

education tests (Watkins & Stanford, 1983; Robbins, 1985). According to a 1987 report from the U.S. Department of Education, more than 20 states now mandate some form of testing for candidates to be admitted into teacher education programs (Office of Education Research and Improvement, 1987, p. 14).

Given the recent evolution in standards for admission of students into teacher-education programs, it seems likely that in most institutions, admission will be based on some combination of previous academic record, test scores, and personal characteristics. Before instituting any new evaluation mechanisms for admission into teacher-education programs, institutions would be well advised to scrutinize closely whether new forms of evaluation are providing significantly new or different information beyond that which is normally gathered in the process of college admission and in the early years of college prior to formal admission into teacher-education programs. History clearly shows that market forces, such as the numbers of students applying to teacher-education programs, surely will condition how any such standards are applied. Thus, the most effective means for improving the qualifications of teacher-education students likely lies not with instituting new forms of evaluation, but with efforts to enlarge the size of the pool of applicants to teacher-education programs.

IV. Requirements for Continuation

Once students are admitted into a teacher-education program, they must meet certain standards in order to continue in the course of study chosen. These are generally defined in terms of patterns of courses which must be taken and grades that must be obtained in those courses.

Over the last half century — up until very recently — there has been a tremendous increase in the professional education courses that preservice teachers were required to take as part of their college careers. In the late 1960s, states required between 15 and 36 semester hours of professional education course work for elementary teachers and between 12 and 29 for secondary teachers (Stinnett, 1967, p. 249). This is equivalent to some 10% to 30% of the semester hours typically required for an undergraduate degree.

Recently, however, a variety of proposals have called for reforming the undergraduate education of would-be teachers so that undergraduate course work in education be reduced, if not altogether abolished. The Holmes Group (1986) for instance proposed that "the undergraduate education major must be abolished in our universities" (p. 14) and the professional course work for teachers take place mainly at the postgraduate level. Similarly, the governors' report on education, *Time for Results* (Alexander, Clinton, & Kean, 1986), called for rebuilding the system of teacher education:

These plans involve new reliance on graduate programs, "clinical" schools that would train teachers while educating young people, the elimination of under-

graduate education majors, and a closer integration of schools and teacher training institutions. They endorse a liberal arts foundation for all teachers. They also would eliminate many course requirements and concentrate on assessment of the knowledge and performance of graduates prior to gaining a state license (p. 38).

What courses one takes in college, however, says little about the manner in which evaluation takes place in those courses. Course evaluation at the college level usually takes the form of grading, with grades being based on some combination of tests, quizzes, written assignments, and perhaps attendance and other assignments. The five-letter system of grading (A, B, C, D, and F) has prevailed in institutions of higher education throughout most of this century and currently is used in over 90% of colleges and universities (Quann, 1984; Levine, 1988). Therefore, it is useful to review briefly some of the literature on college grading (Quann, 1984; Manning, Willingham, Breland, & Associates, 1977).

We treat college grading in general, rather than grading in teacher-training courses alone, for several reasons. First, not much good evidence is available concerning grading in education courses as opposed to college courses more generally. Charges have sporadically been leveled against grading standards in schools of education. Weaver, for instance, mentions a survey of students in "a major eastern university [which] showed that the school of education awarded the highest percentage of A's while being ranked the lowest among the major schools in freshmen average SAT scores," (Weaver, 1983, p. 66), and a study at Temple University (1968) showed that only 2% of students taking an introductory education course received D or F grades, as compared with 30% of students taking introductory liberal arts courses. Nevertheless, we have found very few studies of grading in education courses as opposed to more general studies of college grades.

Second, as previously recounted, certification requirements for education majors mandate less than a majority of semester hours in professional education courses as opposed to more general college course work. Thus, whatever the relative standards of grading in education courses versus other college courses, overall college grade-point averages of prospective teachers generally appear to be dependent only in minor proportion on education course grades.

Third, as also mentioned earlier, in the last ten years or so there have been several initiatives to reduce education course work for prospective teachers at the undergraduate level. Thus, it appears that general college courses and grades in them will be of increasing importance in the preservice evaluation of prospective teachers in the foreseeable future.

While it is not possible here to review thoroughly the literature on college grading, three general points about college grading and teacher evaluation are clear. First is that college grades are increasingly important criteria in terms of which prospective teachers are evaluated. In the late 1970s and early 1980s, there has been a definite trend for states to increase the college grade point

average required for admission into teacher-training programs. As Sandefur (1984) reported,

Although grade point average (GPA) has been a criterion for admission to teacher education programs for many years, there has been a pronounced trend to raise the GPA [requirement] in recent years (p. 5).

In 1979, for example, the Alabama State Board of Education set out new requirements for admission of students into a teacher-education program which included, among other things, that "The student must have a minimum grade point average (GPA) of at least 1.2 on a 3.0 scale on all college work attempted" (Watkins & Stanford, 1983, p. 2). Similarly in the early 1980s, a commission in Georgia recommended that the grade-point average for admission to undergraduate teacher education programs in that state be raised from 2.5 to 2.7 on a 4-point scale (Robbins, 1985, p. 2).

A second general comment about college grading practices is that they have been regularly criticized through most of this century (see, for example, Kirschenbaum, Simon, & Napier, 1971). Dressel (1983) encapsulated much of the criticism of grading as a means of evaluation in his quip that a grade is "an inadequate report of an inaccurate judgment by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of an indefinite material" (p. 12). Grades awarded by instructors have been shown to vary from one occasion to another even when the same work was being graded. Different instructors may award different grades to the same papers, and standards of grading appear to vary with the general level of ability of students enrolled in a course. Moreover, grading standards within the same institution can change markedly over time. In a period of college grade inflation from the 1960s to the mid-1970s, the percentage of A grades given in college courses doubled in many institutions and a national survey revealed that more than three quarters of college faculty surveyed identified grade inflation as a serious problem (Levine, 1988, pp. 102-103).

Quite apart from their deficiencies in terms of validity and reliability, grades have also been criticized for their effects on student motivation and attitudes toward learning. In a national survey, 47% of students reported that they had dropped (or only audited) a course because they were afraid of receiving a poor grade in it, and over half of the responding students acknowledged that they had cheated on one or more occasions because they wanted better grades (Milton, Pollio, & Eison, 1986, pp. 66-67).

Above and beyond these problems, the manner in which grade-point averages are calculated across courses is fraught with problems. Among the ways in which institutions of higher education differ in calculating GPAs are the numerical scale used, the treatment of grades in courses repeated, and differential weighting of remedial and honors courses. Because of such problems with college

grading and the calculation of GPA's, Milton, Pollio, and Eison (1986) proposed abolishing the college GPA:

... it is a useless and misleading statistic for either teaching or research purposes. If administrators or researchers feel the need for an overall summary of student college learning, we should redesign transcripts so they show patterns in a student's academic career; let us never condone or condemn a student on the basis of a single number, artificially significant to two decimal places (Milton, Pollio & Eison, 1986, p. 203).

In addition to these general criticisms of college grading, available evidence indicates that how well a student does in college as measured by grade-point average has little power to predict how well the individual will succeed as a teacher. In a summary of studies stretching over 40 years, for example, A. S. Barr and colleagues (1961) found that none of some 180 measures of teacher characteristics investigated, including college grades, had much power to predict teacher effectiveness, whether it was measured via ratings of school supervisory personnel, teacher educators, teacher peers, pupils, or by pupil gains on tests. In a summary relating various kinds of college grades to different indicators of teacher quality, Barr et al. showed that most correlations were small in magnitude and not statistically significant. Among some 60 such correlations between various kinds of college grades and various presumed indicators of teacher quality, about a quarter of the associations were negative and around 80% of the correlations were less than 0.30. About the only consistently positive and significant correlations between grades and an indicator of teacher quality were those between upper-level college grades and practice teaching grades (Barr et al. 1961, p. 35) — a not terribly surprising association.

The studies summarized by Barr et al. (1961) included one which sought to look at the effects on teacher quality of raising the GPA required for entry into teacher-education programs:

Stoelting (1955) found for University of Wisconsin graduates that if the minimum grade-point average were increased from 1.3, now employed, to 1.5, formerly employed, the higher minimum would screen out 13 of 24 who were judged to be less than average in teaching ability but at the same time eliminate 31 who were average, 7 who rated above average, and 1 who rated superior (Barr et al., 1961, p. 144).

Many more recent studies have examined evidence concerning the relationship between college grades of teachers and success in teaching, with the latter defined in a variety of different ways. While the association between grades and teaching success has more often been found to be positive than negative, the degree of association usually has been found to be not statistically significant (Barnes, Blaisdell, & Hall, 1977; Bode, 1972; Dawson, 1972; Fratiani, 1979;

Pratt, 1984, 1987; Quirk, Witten, & Weinberg, 1973). In other words, despite recent efforts to improve standards of teacher education by increasing both the entry and exit grade point average requirements of teacher-education programs, a fairly long history of research indicates that college grades generally have little power to predict teacher quality.

This finding is surely attributable partly to the difficulty of measuring teacher quality (invalidity and unreliability in a criterion measure will lessen the observed correlation between a predictor and the measure of the criterion of interest). Nevertheless, that college grades have little power to predict teaching success is hardly surprising in light of the broader finding that college grades have little power in general to predict success in adult life after college, whether such success is measured in terms of ratings of on-the-job performance, earnings, or occupational status (see for example, Hoyt, 1985). Indirectly grades do, of course, have an influence on adult success, as some minimum defined in terms of grades is universally required for college graduation, and college completion has been shown to confer an earning advantage on adult workers (see for example, Jencks et al., 1979).

As course grading is perhaps the most common form of evaluation of preservice teachers and college grading has been widely shown to have numerous deficiencies, efforts to improve grading standards would seem to be one of the most promising avenues for improving the evaluation of preservice teachers. Among the suggestions for improving college grading practices (Milton, Pollio & Eison, 1986, chapter 9; Levine, 1988, chapter 4) have been:

Giving college professors more guidance in appropriate grading standards and in building better classroom tests and other assessment devices from which they usually derive grades;

Administrative review or even enforcement of constraints on the distribution of grades that professors may award;

Supplementing letter grades with written evaluations or other more specific evaluations of student strengths and weaknesses.

As previously noted Milton, Pollio, and Eison (1986) even advocated the abolishment of the college grade-point average altogether, but because letter grades and the GPA are well entrenched in the policies and practices of institutions of higher education (and implicitly required, as noted above, in some state policies regarding the preservice education of teachers), such action seems highly unlikely. More practical advice for any institution preparing teachers would be to examine closely procedures for calculation of undergraduate GPAs, to ascertain whether such procedures are not only fair and reasonable but also guide students in the intended directions with regard to both course taking and work in courses.

In recounting such advice, it should be acknowledged that changing college grading practices is surely not easy. For one thing, traditions of academic

freedom make many college professors resistant to advice as to how to run and grade their courses. Less obviously, college grading practices have been fairly clearly shown to be an adaptive mechanism, with grading standards in different disciplines influenced by the caliber of students in the field (Milton, Pollio, & Eison, 1986, chapter 8; Goldman & Hewitt, 1975). In other words, a program or professor with a shortage of students is much less likely to flunk students than the program or professor with standing room only. This is of course simply another wrinkle on the market forces dynamic already mentioned. Like college admission standards, grading standards appear to be clearly influenced by the supply of students.

V. Practice Teaching

If college grades in general do not seem to have much power to predict teaching success, a natural next question is whether success in practice teaching is a better means of evaluating preservice teachers. Practice teaching is, after all, a virtually universal requirement for teacher certification. Practice teaching or some other form of actual teaching experience satisfactorily completed has, in the last 30 years, come to be a virtually universal requirement for permanent teacher certification in all states of the country (Burks, 1987; Woellner, 1972, 1977, 1982; Woellner & Wood, 1945, 1963, 1967).

Moreover, as noted above, some recent reform proposals for teacher education would place more emphasis on clinical teaching experiences as a means of preparing prospective teachers than on college studies in education courses.

Nevertheless, available evidence on practice teaching suggests that prospective teachers' performance in practice teaching is not a good predictor of subsequent teaching performance. Here, for example, is Pratt's (1984) summary of relevant evidence:

Browning (1988) found a correlation of $-.03$ between practice teaching ratings by education professors and ratings of the same teachers in their first years of teaching by supervisors; the agreement between in-school supervisors in practice teaching and in the first year of teaching was $.23$. Jenkins (1978) on the other hand, found a correlation between practice teaching and first-year principal's rating significant at $.01$. The inconsistent predictive value of practice teaching scores calls into question much of the research on prediction of teaching ability which uses practice teaching as the dependent variable (Pratt, 1984, p. 4).

Additionally, the summary by Barr et al. (1961), already mentioned, showed no indication that grades in practice teaching were a better predictor of subsequent inservice ratings of teacher quality than were grades in other college courses.

Part of the reason for this may well be that standards for grading in practice teaching are likely even more varied than grading standards in college courses more generally. While we have not located any systematic studies of methods of evaluation of practice teaching generally, we have heard anecdotal accounts to suggest that once a student has progressed through a teacher-education program to the point of receiving a practice teaching assignment, college supervisors are reluctant to fail the practice teacher. After all, a student who fails practice teaching implicitly reflects badly on the previous teacher-education course work that the student has completed. Again according to anecdotal reports, it appears that a student who performs badly in a practice teaching assignment is more likely to be counseled (or self-selected) out of pursuing a teaching career than to be evaluated as failing a practice teaching assignment.

Nevertheless, though means of evaluating practice teaching appear to have been little studied in the past, it seems likely that standards of performance in practice or clinical teaching experience will be receiving increasing attention in the future. For, as previously recounted, many of the recent teacher-education reform initiatives in the last ten years have been focusing increasing attention on the provision of clinical teaching experiences of prospective teachers.

Summary and Conclusion

Graduation from college has become a minimal requirement for a person to become a teacher at the elementary or secondary level. Nevertheless, less than one quarter of the adult population aged 25 to 34 years has completed four or more years of college (Stern, 1988, Vol. 2, p. 17). These facts clearly show that from a broad labor market perspective, the screening and evaluation of people prior to completion of college have more impact on who will become elementary and secondary school teachers than do the last few steps (such as teacher-competency testing) in the teacher certification process.

While admissions into some colleges remains highly selective, admissions standards for colleges and universities in the United States vary markedly. Available evidence suggests that with the decline in size of college age population from the 1960s into the 1980s, gaining admission into college is now easier than it was 20 years ago. It also appears that as the proportion of college freshmen expressing a preference for a career in the elementary or secondary teaching has declined from the 1960s to the 1980s, standards for entry into teacher-education programs may have declined. After widespread criticism of admissions standards for entry into teacher-education programs beginning in the late 1970s, a variety of proposals have been made to upgrade such standards in the future, for example by instituting basic skills assessments or higher grade-point average requirements of teacher education candidates.

Once students are admitted into teacher-education programs, they typically must meet requirements defined in terms of courses that must be taken and grade point averages that must be attained. Up until recently, teacher-education majors

typically took between 10% and 30% of their total number of college course hours in professional education course work. Recently, however, a number of proposals have been advanced to reduce the number of teacher education courses for education majors and to increase the proportion of more general liberal arts courses.

College course work is almost universally evaluated in terms of the five-letter system of grading. Nevertheless, letter grading practices of college professors have long been criticized for their variability, unreliability, and invalidity. Additionally, institutional practices for calculating grade-point averages are sufficiently varied so as to have led to severe criticisms of the college GPA as a meaningful measure of students' performance. Also, research has shown college GPA to be an extremely weak predictor of ratings of teacher quality (and other measures of success in adult life after college). While a variety of strategies have been suggested for improving college grading practices, traditions of academic freedom appear to have made grading practices somewhat impervious to systematic change. Nevertheless, college grading practices do appear to change in response to market forces, such as the number and caliber of students entering particular institutions, fields, and courses.

In addition to traditional college course work, preservice teachers are required to pass a practice teaching assignment in order to attain a permanent teaching certificate in all 50 states in the country. While little systematic evidence has been located on the exact bases by which students are evaluated during practice teaching, several studies have shown that grades in practice teaching have no more power than college grades, generally, to predict ratings of teachers by principals or other supervisors. Anecdotal accounts also indicate that relatively few students actually flunk practice teaching assignments, but instead tend to be counseled or self-selected out of teaching careers as a result of unsuccessful practice teaching experiences.

A variety of strategies have been proposed for improving the screening and evaluation of preservice teachers, including changing the requirements for admission into college and/or teacher-education programs, instituting entrance and exit exams, and improving the grading practices of college professors. While so-called teacher-competency exams have not been systematically reviewed in this chapter, it appears from the history of teacher education in the United States and from a variety of literature on the preservice evaluation of teachers that while such proposals may lead to marginal improvements, the caliber of students in preservice education and the manner in which they are evaluated is likely to be much more strongly affected by market forces concerning the number and quality of students intent on going into teaching. In short, it seems clear from evidence concerning both the teacher labor market and college grading practices that formal means of evaluating prospective teachers offer little leverage for improving quality of the teacher labor force, as long as the supply of prospective teachers does not significantly exceed demand. This suggests that one of the most promising strategies for improving the caliber of prospective teachers, and

indirectly their evaluation, is simply the recruitment of larger numbers of students into teaching careers. As for evaluation per se, given widely documented weaknesses in college grading practices and in teacher testing, a reasonable strategy appears to be avoiding overreliance on either one of these two main forms of preservice evaluation and to employ multiple lines of evidence, analogous to what Peterson (1988, chapter 6, this volume) has advocated with regard to the inservice evaluation of teachers.

References

- Alexander L. Clinton, B. & Kean, T. (1986). *Time for results: The governors' 1991 report on education*. Washington, DC: National Governors Association.
- Astin, A., Green, K., Korn, W. & Schalit, M. (1986). *The American freshmen national norms for fall 1986*. Los Angeles, CA: University of California, Los Angeles, Higher Education Research Institute.
- Barnes, C., Blaisdell, L., & Hall, S. (1977). Darts or criteria? How do we choose teacher education students? *California Journal of Teacher Education*, 4(3) 63-73.
- Barr, A.S., Worcester, D.A., Abell, A., Beecher, C., Jenson, L.E., Peronto, A.L., Ringness, T.A., & Schmidt, J., Jr. (1961). *Wisconsin studies on the measurement and prediction of teacher effectiveness*. Madison, WI: Denbar Publications.
- Bode, J.C. (1972). *Criteria for determining probable success of future teachers of vocational agriculture*. Doctoral dissertation, Iowa State University.
- Browning, R.C. (1968). Validity of reference ratings from previous employers. *Personnel Psychology*, 21(3), 389-393.
- Burks, M.P. (1987). *Requirements for certification of teachers and administrators*. Chicago: University of Chicago Press.
- Dawson, D.A. (1972). Educational quality indices. *Alberta Journal of Educational Research*, 18, 35-44.
- Digest of Education Statistics*, 1988. Washington, DC: U.S. Government Printing Office.
- Dressel, P. (1983). Grades: One more tilt at the windmill. In A. W. Chickering (Ed.), *AAHE Bulletin*, 35(8), 10-13. Cited in Milton, Pollio, & Eison (1986) p. 23.
- Evenden, E. S. (1933). Summary and interpretations. *National survey of the education of teachers*, (6). U.S. Office of Education, Federal Security Agency, Bulletin 1933, No. 10. Washington, DC: U.S. Government Printing Office.
- Feistritzer, C.E. (1984). *The making of a teacher: A report on teacher education and certification*. Washington DC: National Center for Educational Information.
- Fratiani, J. (1979). *Prediction of beginning teaching success from personal and program variables*. Doctoral dissertation. University of Wisconsin.
- Goldman, Roy D. & Hewitt, Barbara Newlin, (1975). Adaptation-level as an explanation for differential standards in college grading. *Journal of Educational Measurement*, 12(3) 149-61.
- Grant, W.V. & Snyder, T.D. (1983). *Digest of Education Statistics, 1983-84*. Washington, DC: U.S. Government Printing Office.
- Haney, W. (1984). Testing reasoning and reasoning about testing. *Review of Educational Research*, 54(4), 597-654.
- Haney, W. Madaus, G. & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. In E. Rothkopf (Ed.), *Review of Research in Education* (pp. 169-238). Washington DC: American Educational Research Association.
- Hodges, H.P. & Nash, A. (1982, Mar. 5). *Should teacher training institutions be concerned about the reading and writing skills of future teachers?* Paper presented at the annual meeting of the Washington Organization Reading Development Association for International Reading Association. Seattle, WA. (ERIC ED236 579).
- Holmes Group. (1986). *Tomorrow's teachers*. East Lansing, MI: Author.
- Hoyt, P. Donald. 1965. *The relationship between college grades and adult achievement*. (ACT Research Report No. 7) Iowa City, IA: American College Testing Program.

- Information Please Almanac, 1987. Boston: Houghton Mifflin.
- Ishler, R.E. (1984). Requirements for admission to and graduation from teacher education. *Phi Delta Kappan* (October), 121-122.
- Jencks, C., Bartlett, S., Corcoran, M., Crouse, J., Eaglesfield, D., Jackson, G., Michelland, K., Mueser, P., Olneck, M., Schwartz, J., Ward, S., & Williams, J. (1979). *Who gets ahead: The determinants of economic success in America*. New York: Basic Books.
- Jenkins, H.A. (1978). *The relationship of beginning teacher scores on the national teacher exam and other selected variables to their competency in teaching*. Unpublished doctoral thesis, Mississippi State University.
- Kirschenbaum, H., Simon, S., & Napier, R. (1971). *Wad-ja-get? The grading game in American education*. New York: Hart Publishing.
- Levine, A. (1988). *Handbook on undergraduate curriculum*. San Francisco: Jossey-Bass.
- Manning, W., Willingham, W., Breland, H., & associates (1977). *Selective admissions in higher education*. San Francisco: Jossey-Bass.
- Milton, O., Pollio, H., & Eison, J. (1986). *Making sense of college grades*. San Francisco: Jossey-Bass.
- National Council for Accreditation of Teacher Education. (1985, April). *NCATE redesign*. Washington, DC: NCATE (1919 Pennsylvania Ave. NW, Suite 202).
- Nelli, E. (1984). A research based response to allegations that education students are academically inferior. *Action in Teacher Education Journal*, 6(3), 73-77.
- Office of Educational Research and Improvement (1987). *What's happening in teacher testing?* Washington, DC: U.S. Government Printing Office.
- Plisko, V.W. (Ed.) (1983). *The condition of education*. Washington, DC: U.S. Government Printing Office.
- Pratt, D. (1984, June). *Predicting career success in teaching*. Paper presented at the annual conference of the Canadian Society for the Study of Education, Guelph, Ontario. (ERIC ED 273 603)
- Pratt, D. (1987). Predicting career success in teaching. *Action in Teacher Education Journal*, 8(4) 25-34.
- Quann, C. (1984). *Grades and grading: Historical perspectives and the 1982 AACRO study*. Washington, DC: American Association of Collegiate Registrars and Admissions Officers.
- Quirk, T.J., Witten, B., & Weinberg, S.F. (1973). Review of studies of the concurrent and predictive validity of the national teacher examinations. *Review of Educational Research*, 43(1), 89-113.
- Resnick, D. (1982). History of educational testing. In A. Wigdor and W. Garner (Eds.), *Ability Testing* (Volume 2, pp. 173-194). Washington DC: National Academy Press.
- Robbins, J. (1985). *Comments on grade point average for admission to teacher and administrator preparation programs*. Unpublished paper, College of Education Georgia State University. (ERIC No. ED 256720).
- Sandefur, J.T. (1984). *Standards for admission to teacher education programs. An issue paper prepared for the Minnesota higher education coordinating board*. Unpublished paper, College of Education, Western Kentucky University. (ERIC No. ED251417).
- Sedlack, M. & Schlossman S. (1987). *Review of Research in Education*, 14, 93-131.
- Snyder, T. (Ed.) (1987). *Digest of educational statistics 1987*. Washington DC: U.S. Department of Education, Office of Educational Research and Improvement, Center for Education Statistics.
- Stern, J.D. (Ed.) (1988). *The condition of education. Elementary and secondary education 1988. Volume 1*. Washington, DC: U.S. Government Printing Office.
- Stern, J.D. (Ed.) (1988). *The condition of education. Postsecondary education 1988. Vol. 2*. Washington, DC: U.S. Government Printing Office.
- Stinnett, T.M. (1967). Teacher certification. *Review of Educational Research*, 37 (June), 248-252.
- Stoelting, G.J. (1955). The selection of candidates for teacher education at the University of Wisconsin. *The Journal of Experimental Education*, 24, 115-140.
- Tanner, D. (1986, Feb. Mar.). *Teacher education and academic rigor*. Paper presented at the Annual meeting of the American Association of Colleges for Teacher Education, Chicago, IL. (ERIC ED272 505).
- Temple University. (1968). *Report of the college of education ad hoc committee on grading systems*. (Cited in Kirschenbaum, Simon, & Napier, 1971).
- Watkins, J.F. & Stanford, R.L. (1983, Jan. 30, Feb. 2). *ACT scores and selective admissions: An exploratory look at some one-time data*. Paper presented at the National Conference of Teacher Educators, Orlando, FL. (ERIC No. 230520)

- Weaver, W.T. (1979). In search of quality: The need for talent in teaching. *Phi Delta Kappan*, 63(2), 106-112.
- Weaver, W. T. (1983). *America's teacher quality problem*. New York: Praeger.
- Weaver, W.T. (1984). Solving the teacher quality problem. *Phi Delta Kappan*, 66(2), 108-114.
- Willingham, W. & Breland, H. (1977). The status of selective admissions. In Manning, W. et al., *Selective admissions in higher education*, pp. 65-152. San Francisco: Jossey-Bass.
- Woellner, R.C. & Wood M.A. (1945). *Requirements for certification of teachers and administrators*. Chicago: University of Chicago Press.
- Woellner, E.H. & Wood M.A. (1963). *Requirements for certification of teachers and administrators*. Chicago: University of Chicago Press.
- Woellner, E.H. & Wood M.A. (1967). *Requirements for certification of teachers and administrators*. Chicago: University of Chicago Press.
- Woellner, E.H. (1972). *Requirements for certification of teachers and administrators*. Chicago: University of Chicago Press.
- Woellner, E.H. (1977). *Requirements for certification of teachers and administrators*. Chicago: University of Chicago Press.
- Woellner, E.H. (1982). *Requirements for certification of teachers and administrators*. Chicago: University of Chicago Press.

4. Licensure and Certification of Teachers: An Appraisal

Gary Sykes

Introduction

This chapter considers the standards (i.e., minimum qualifications) associated with the issuance of teaching credentials. Looking across occupations, two sorts of credentials are evident, often confused in terminology and in the public mind. *Licensure* is a process controlled by state governments in the name of public health, safety, and welfare. Licenses issued by state agencies legally restrict practice to those duly licensed. Standards for licensure are minimum qualifications designed to protect the public from incompetence. States grant licenses to practice a wide variety of trades and professions — everything from cosmetology and barbering to lawyering and doctoring. The number of occupations so regulated varies from state to state, but the total approaches 800 overall (Shimberg, 1985).

Certification is a process controlled by particular occupational groups that typically signifies special or advanced competence in a field of practice. Certification is a voluntary process that does not involve restriction on the right to practice. Certification standards are most highly developed in the health field where physicians, osteopaths, dentists, nurses, pharmacists, and others can become certified in a variety of specialties.

These two terms, however, are not used uniformly across occupations, giving rise to much confusion. In the teaching field, *certification* refers to state requirements for practice, rather than the more widely used term, which is *licensure*. In analyzing the case of teaching, I will adhere to education usage, which will be familiar to most readers of this volume.

AUTHOR'S NOTE: This paper was written in conjunction with a project sponsored by the Center for Policy Research in Education, and funded by the U.S. Department of Education, Contract No. OERI-0086-90011. The views expressed in this paper are the author's and do not necessarily reflect the view of the U.S. Department of Education or the institutional partners for the Center for Policy Research in Education.

Licensure constitutes government restriction of free trade in a service market and has a twofold rationale. First, services such as education have social effects that extend beyond individual transactions. To ensure an acceptable level of quality in these social benefits, governments claim the right to intervene in such service markets. Second, under market conditions where buyers do not process the knowledge and skill to evaluate complex services, government intervention is warranted in the name of consumer protection. Finally, in the case of a public service such as education, the establishment of tenure after a short period precludes hiring and firing in response to client demands.

By contrast, the rationale for certification is the articulation of professional standards by an occupational group seeking the public's trust. Certification represents advanced standing in a field, attained through training and demonstration of competence. Whereas licensure is a mandatory process that restricts practice, certification is a voluntary process that confers selective benefits, including, for example, professional status, pay, advanced positions, hospital privileges, reduced insurance rates, and others.

In practice, these processes are complexly interrelated. In medicine, the National Board of Medical Examiners (NBME) set out in 1915 to create a voluntary standard that would influence licensure. Gradually, over the next 40 years, this occurred as state after state came to accept the NBME's tripartite requirement of graduation from an accredited medical school, a three-part, staged examination, and completion of a supervised, year-long internship. In other fields, licensure itself constitutes a form of advanced standing, for it is possible to practice on a limited basis without a full license. Engineers, architects, and accountants may practice in the field but on a restricted basis. To become a full professional — a registered engineer, a registered architect, or a certified public accountant — involves meeting requirements of training, experience, and examination established by the profession but legally sanctioned by state boards that are linked in national federations.

The modern structure of standard-setting took root during the Progressive Era, and has evolved steadily toward increased formality and stringency. To appraise licensure and certification in teaching, this chapter advances a systemic framework that accounts for three broad and interacting effects of standard-setting: the generation of teacher supply within the political economy of the educational system; the construction of categories defining the credential system; and the definition of teaching competence as a normative, conceptual, and empirical invention. Sections to follow take up each topic in turn, with a concluding appraisal of options for the future.

Generating Teacher Supply

A variety of factors interact in generating the supply of teachers, most obviously teacher salaries compared with wages of other occupations competing for

college talent. Credentialing processes are commonly thought to restrict, not generate, supply by establishing minimum qualifications for entry, so it seems peculiar to credit credentialing with a role in producing the supply of teachers. But from a systemic perspective, generating teacher supply is a critical function of credentialing. To understand this perspective requires a brief excursion into the political economy of the professions.

Standards established through state regulation of occupations create monopolies of competence, but in no cases are those perfect monopolies. In all professional labor markets, clearing mechanisms rely on the joint fluctuation of wages and standards. The response to shortages typically includes a combination of wage increases with a relaxation or redefinition of standards. The balance between these two mechanisms varies from field to field, however, as do the particular means for defining, differentiating, or degrading the standards.

In nursing, for example, World War II produced an enormous nurse shortage, in response to which a new category of nurse — the practical or vocational nurse — was created to supplement the more highly trained registered nurse. Today, both categories of practitioner exist, each with its own training programs, practice laws, and institutional roles. In medicine, standard-setting restricted physician supply and drove up wages, but other expedients took shape as well to meet demand for medical services. A variety of irregular practitioners — osteopaths, chiropractors, midwives, faith healers — survived elite medicine's efforts to eliminate competition, and they continue to serve many communities. And, supplementary personnel such as physician's assistants have arisen in the field to supply medical services to underserved areas. In others fields, notably architecture, accounting, and engineering, credentialing laws protect the title, not the practice. This wrinkle on licensure means that individuals may practice on a restricted basis, but may not identify or advertise themselves as certified or registered. Much routine work in these fields is carried out by such unlicensed practitioners.

Excluding the military establishment, education is the nation's largest public monopoly. To staff our schools has required upwards of one fifth of the college labor market historically. In most states, educational expenditures are the largest item in the budget, with teacher salaries accounting for a sizable share of the total. Teacher salaries are not market-sensitive; they are subject to the politics of the budgetary process at state and local levels, and lags in the teacher labor market are the rule, not the exception. Furthermore, teaching has undergone little differentiation over the years. Teaching's circumstances, then, tip the balance toward flexibility in standards as a means of generating supply, and five mechanisms are evident.

States may adjust licensure test cut scores downward to increase supply; they may grant a variety of substandard credentials that circumvent requirements for full credentials; and they may sanction alternate route programs to recruit non-traditional candidates by reducing opportunity and other costs of entry. School districts contribute two additional mechanisms. They may assign teachers to

classes outside their area of licensure on a full or part-time basis. And they may increase class size to accommodate teacher shortages. From a public policy perspective, an appraisal of any state's teacher credentialing *system* requires attention to these processes in addition to the licensure standards and procedures.

Methods for setting cut scores on teacher licensure tests generally involve assembling a panel of expert reviewers who make judgments about expected performance on individual test items against the hypothetical standard of the "minimally competent teacher." These judgments then are aggregated and analyzed in a variety of ways to establish an overall cut score for the test (Haney, Madaus, & Kreitzer, 1987). States then use this "study score" to set a passing (cut) score, with most states adjusting the cut score downward to minimize false negatives. For example, one study notes that in 12 of 13 states using the National Teacher Examination, the mean adjustment amounted to two standard errors of measurement downward on the Communications Skills, General Knowledge, and Professional Knowledge tests (Office of Educational Research and Improvement, 1987). The authors further note that,

With these downwardly adjusted passing scores, the actual number of items one needs to answer correctly in order to pass the examinations is relatively low. The 1982 Professional Knowledge Test, for example, is comprised of 104 questions. The average passing score set by states using the test was 642. To obtain this score, one only needs to answer 47 of the 104 items correctly. Passing scores range from 35 items in the state with the lowest passing score to 53 items in the state with the highest passing score. The test is scored on the basis of the number of correct answers without subtracting points for incorrect answers or omitted questions. Given the statistical probability, one could expect to answer 21 questions correctly by randomly marking the answer sheet. Candidates also have multiple opportunities to take the examination (p. 37).

Passing rates ranged from a low of 79% in New York to a high of 94% in Kansas, with an average of 87% among these states. Had study scores rather than adjusted scores been used, pass rates would have been 12% to 20% lower on each of the three tests, affecting as many as 8,000 candidates in these 13 states.

Given the leniency of the standard, these figures would be troublesome, but the validity of the test is suspect and the meaning of the cut score questionable. There have been few carefully conducted validity studies of teacher tests, so there is no basis on which to establish cut scores in terms of their correspondence with independent criteria of competence or qualifications. States seek to avoid two evils — admitting unqualified applicants and rejecting qualified ones — but the lesser of these evils appears to be the former. Given a highly fallible indicator of competence — a paper-and-pencil test — most states adjust the cut score to minimize the latter error, and not incidentally enhance the potential supply of new teachers.

The second state expedient for generating teacher supply is the resort to emergency credentials. State requirements for substandard credentials are bewildering in their variety, but roughly one half the states issue day-to-day credentials for short-term substitute teaching of up to one year's duration, with renewal possible in the majority of states.

Data on the nature and issuance of substandard credentials are sketchy, but one tabulation provides a partial indication (Office of Educational Research and Improvement, 1987, p. 45). In 24 of 37 states reporting, less than 2% of teachers were on emergency credentials; in 6 states, between 2% and 4%; and in 7 states, more than 4% (Georgia, for example, indicates 17.6%, Florida 11.6%, and both West Virginia and New Mexico over 5%). A national teacher survey in 1986 indicated that 8.4% of respondents held an emergency or provisional certificate (National Education Association, 1987, p. 20).

A more recent expedient is the creation of state-mandated alternate route programs into teaching. Some 20 states have established such programs, and they are under study in another eleven states (Roth & Lutz, 1986). These programs provide means for individuals to enter teaching without meeting the full set of licensure requirements. They aim to recruit capable college graduates by minimizing entry requirements. In some cases, notably New Jersey, alternate route programs are intended to replace emergency credentials. Elsewhere, they coexist with substandard credentials.

A fourth expedient to meet demand, the misassignment of teachers, is yet more dubious on quality grounds. Nineteen states allow employed/credentialed personnel to teach outside their legal authorization for limited periods of time. In other states, the law does not sanction misassignment, and a variety of penalties and reporting mechanisms have been set up to control this process.

A 1986 national survey revealed that 17% of teachers report they are misassigned, with over 10% indicating misassignment for more than half of their teaching day (National Education Association, 1987, p. 29). The evidence suggests that misassignment is concentrated at the middle school and junior high school level, which one report characterizes as a no-man's-land, a mishmash of teachers with elementary and secondary certification (Counsel for Basic Education, 1985). Shortage areas — mathematics, science, bilingual education — are likely to include high proportions of misassigned teachers, and rural schools with small staffs typically must stretch their faculties across multiple assignments in order to offer a broad array of courses.

A final source of administrative flexibility is class size. If there are 150 students enrolled in ninth-grade English, then they can be grouped into 5 classes of 30 students, 6 classes of 25, or 7 classes of 21-22. Class size is an issue in collective bargaining, and contracts set limits in many locales, but necessity often dictates exceeding such limits. In this case, local administrative action serves to reduce demand for teachers. As with teacher misassignment, evidence indicates a steady reduction in class size over the past quarter century. In 1986, elementary classes averaged 24 pupils (compared to an average of 29 pupils in

1961), and secondary classes averaged 26 (in 1961, the average was 27) (NEA, 1987, pp. 33-34).

Teaching is so large and varied an enterprise in the United States that the ideal of a perfect supply of fully qualified professionals to serve every community is unattainable. Every state and district makes some use of these expedients for generating supply, but the long-term trend is clear. Standards have risen gradually, and the certification system has steadily improved the supply of teachers made available to the schools.

Constructing the Categories of Competence

If one function of the teacher credential system is to generate supply, another is to construct the supply in terms of officially designated categories. These two functions interact. The broader the credential category, the more flexibility employers have in hiring and placing teachers. Broad-based credentials by definition reduce recourse to misassignment and so lend the appearance of high standards. A state may issue a certificate in "secondary science" that entitles its holder to teach biology, earth science, chemistry, or physics, but it is doubtful anyone has mastered this range of subjects.

The credential system is a primary means through which professions introduce specialization into practice, a trend most evident in the health field, where certification came to be associated with specialized medicine. Throughout the middle decades of the twentieth century, the central conflict in the field involved general versus special practitioners, but the accelerating accumulation of scientific knowledge and technology pressed toward specialization (Stevens, 1971).

Medicine's structure became specialized but managed to embrace general practice as well via the introduction of *horizontal* specialties alongside the *vertical* specialties. The former involve general areas of practice, where the physician must exercise broad competence rather than acquire specialized knowledge in a particular area (e.g., brain surgery). The horizontal specialties include family practice and emergency medicine, the former created specifically to accommodate yesteryear's G.P. More than 20 additional specialties and numerous subspecialties constitute the remainder of medicine's categorical structure, with new areas continuing to emerge (e.g., gerontology, nuclear medicine) in response to social and technical developments.

Categories of competence, then, constitute a crucial and dynamic element in any system of specialized credentials (the law is a contrasting case where specialization in practice has not produced specialization either in formal training or in the credential structure). Again, the case of teaching is particularly complex because of the range of subjects taught and the range of grade configurations composing public schools. Teaching also has its horizontal and vertical specialties. Elementary teaching requires broad knowledge of a number of subjects and their interrelationships, while secondary teaching is partially

specialized (but contrast the teaching of physics and social studies; the latter requires knowledge of history, economics, political science, geography, and perhaps sociology and anthropology).

States vary considerably in their credential categories along three dimensions: breadth, level, and duration. In 1986, at least 26 states issued *semibroad* certificates, that allowed teaching in such areas as social studies, language arts, or science (Roth, 1986). However, many states have turned to specific credentialing, usually involving requirements that a teacher must major or minor or have taken a specified number of credits in the subject area covered. The number of such specific endorsements also ranges widely; Nebraska has 20, Massachusetts has 52.

The second dimension is the number of grade levels covered by the credential. In most states the early elementary grades and upper secondary serve as unambiguous targets, but the middle grades create problems. Junior high schools and middle schools feature a wide array of grade configurations staffed with teachers holding either elementary or secondary credentials. Some states have special requirements for elementary credential holders who teach departmentalized courses in grades 7-9; others do not differentiate their elementary credentials. Most states require a match between secondary credentials and the subject taught in grades 7-9, but some states issue substandard credentials for these levels. And, a small but growing number of states have established a separate middle school credential (NASDTEC, 1988).

States also vary in authorizing the duration of credentials. In the past, lifelong credentials were the rule, a practice still followed by 16 states. Thirty-two states, though, have established requirements to renew credentials based on continuing education or continuous years of experience, and a trend is toward more stringent and tightly focused course requirements. Three states — Arkansas, Georgia, and Texas — also test teachers for recertification. The tests focus on basic skills (and on subject matter knowledge in the Georgia case) and can result in revocation of the license to teach.

The long-term secular trend has been to reduce the duration of credentials and to focus the categories more sharply on both subject areas and a limited number of grade levels. This increasing specialization aims to improve the quality of the teacher work force by ensuring a better match between qualifications and assignments. But this trend also reduces staffing flexibility and so countertrends are evident as well in response to pressures from local communities that cannot generate specialist supply.

Most fields rely on scientific rationales to justify specialist credentials; specialization is depicted as a natural outgrowth of knowledge accumulation and technological development. But in all fields, interest group and movement politics play a substantial role in shaping categories of competence. For example, in the teaching field, categories of special education teacher illustrate the interplay of quasiscientific and political factors in shaping the credential system.

In 1988, states issued from 1 to 12 credentials, with an average of 7. In some cases, special needs reside in a physical condition such as deafness or blindness that self-evidently requires special training and qualifications for teachers. But the case of "specific learning disabilities" (SLD) deserves scrutiny. This category has produced a number of specific and generic titles in the states and is recognized as a special-needs area. In fact, 42% of the 4.3 million children formally identified as educationally handicapped are categorized as learning disabled. The term has become a catch-all for children's difficulties with language, spelling, mathematics, and especially reading, where it is synonymous with the older term *dyslexia*.

Specialists attribute learning disabilities to unspecified neurological dysfunctions within the individual, but the specific evidence supporting the etiology of this disorder is highly suspect and evidence suggests that many schools use the LD category as a convenient means of referring children who are difficult to teach out of the regular classrooms. A small number of children do have genuine dysfunctions that impede perceptual and word processing abilities, but the best evidence indicates that this problem has been substantially overgeneralized (see Coles, 1987). In many cases, the difference between "low achievers" and the "learning disabled" appears in the beholder's eye. Conditions of teaching and learning in the schools, not brain malfunctions, are as likely an explanation for learning disabilities.

Categories of competence are social constructions that define the official supply of practitioners and structure the services provided to clients. The credential system is a contested terrain because it directs the allocation and distribution of human resources, confers benefits, and shapes professional practice. Increasing differentiation of practice in the form of specialist credentials tends to tighten the connection between qualifications and assignments, yielding higher quality service in theory. But as the case of learning disabilities makes clear, specialization can produce an equivocal outcome.

Inventing Conceptions of Quality

If the credential system serves important functions related to the supply of teachers, it also defines the knowledge base for practice. The processes of licensure and certification have served as one important locus for the representation and legitimation of professional knowledge in such fields as medicine, law, nursing, architecture, and accounting, to name a few. In these fields, the profession has worked closely with state practice boards to define and refine assessment practices and standards representing the knowledge base.

In most fields, the unevenness of knowledge and disputes among practitioners have complicated efforts to rationalize practice around a common, codified base of knowledge, but the ambiguities of competence in teaching are particularly striking. Little dispute exists that teachers ought to be liberally educated and

should command knowledge of the subjects they teach. However, beyond this basic agreement, consensus breaks down on what individuals need to know and be able to do in order to teach effectively. Teaching, after all, is a ubiquitous activity in our society. It takes place in public and private schools, kindergartens and universities, summer camps and Sunday schools, on television and in the family. Nearly every adult in our society has spent thousands of hours observing teaching. A great deal of this teaching goes unregulated by the state. Teachers within some of our most prestigious institutions of learning — Harvard to Hotchkiss — require no special credentials in order to practice. Widespread, common-sense notions regard teaching as drawing on ordinary knowledge and human qualities, not specialized expertise and codified bodies of technical knowledge (see, for example, Jackson, 1986; and Buchmann, 1987).

Over against this perspective on teaching may be juxtaposed accounts of teaching as complex and subtle work requiring considerable expertise. These accounts emphasize the growing body of technical knowledge that is steadily accumulating around generic teaching competencies, classroom management, effective instructional practices, teaching strategies in the content areas, the nature and development of human intelligence, cultural and social determinants of learning, and other topics. They further emphasize the interactive and adaptive nature of teaching work that requires decision making, judgment, reflection, and the sensitive management of multiple dilemmas of practice. They stress a professional definition of the teaching role that accords teachers broad responsibility for curriculum development and decision making; coordination and collaboration with colleagues; involvement with parents and others in the community; participation in school-wide improvement efforts and in experimentation with new instructional arrangements; and practice in light of the professional ethics of teaching.

These two depictions bracket the range of opinion on teaching, and underlie the currents and cross-currents of change in licensure and certification policy. A number of trends are evident. One involves extension of the clinical portions of teacher education, including more practice teaching, supervised intern and induction programs for beginning teachers, and alternate programs that rely on apprenticeship-like arrangements in schools as the best form of preparation. Practical experience — learning on the job — is granted primacy in these requirements, together with some mediation via supervision.

A second trend favors increased emphasis on subject matter knowledge, especially for secondary teachers. Many states are increasing requirements for liberal studies and for subject majors and minors. Coupled with this trend is a third, the deemphasis of education course work, including both foundations and methods courses. In Florida, for example, a 1985 law required that prospective secondary teachers must earn 30 hours in a subject field outside a department, school, or college of education. Other states (e.g., Texas, New Jersey) have placed limits on the allowable number of education credits and courses for a state-approved degree program.

A fourth trend draws on the research on teaching to establish behaviorally oriented, generic teaching competencies, often embodied in classroom observation systems. A number of states have invested in these on-the-job performance assessments, including Florida, Georgia, North and South Carolina, Virginia and Kentucky, with others likely to follow.

These trends have not escaped criticism and the invention of quality standards in teaching continues to be deeply problematic. The emphasis on increased subject matter knowledge and on strengthening and extending the clinical portions of teacher education have secured the broadest approval and the least criticism. However, the connection between subject matter knowledge represented in university course work and the employment of subject matter knowledge in teaching is unclear. Presumably a mathematics major who takes upper division courses in calculus, set theory, topology, and other advanced topics has a better grounding for teaching ninth-grade algebra or tenth-grade geometry than a student with less in-depth knowledge, but this conclusion is not self-evident and enjoys no empirical support. Furthermore, the extension of clinical experiences does not alone constitute a more effective system of learning to teach. The arrangements for and the quality of supervision, together with the means to integrate practice teaching with structured opportunities to reflect on and learn from practice, are critical. These qualitative aspects of teacher preparation have not received much attention within state policy.

The brunt of criticism, however, has fallen on efforts to define professional knowledge and skill in teaching. Over the years, the Educational Testing Service has gradually expanded and reoriented the professional knowledge component of the National Teacher Examination (NTE), but this section has come under continuous criticism. To a variety of observers over the years, the items have appeared vague and ambiguous, the correct answers often questionable, the research base thin, and the overall sampling domain ill-defined (see Haney et al. [1987] for review of these critiques; see Darling-Hammond [1986] for a contemporary critique). Furthermore, reviews of NTE validity studies have consistently called this examination into question (see Quirk, Witten, & Weinberg, 1973; and Ayers & Qualls, 1979).

Equally critical have been reviews of the various competency-based, performance-oriented approaches to teaching assessment. These critiques, emphasize a range of conceptual and methodological shortcomings, including their attention to teacher behavior alone, their tendency to imply a single, generic model of good teaching that leaves subject matter (among other things) out of the account, and their inability to account for contextual factors that influence teacher actions and decisions (see, for example, Macmillan & Pendlebury, 1985; Shulman, 1986; and Tyson-Bernstein, 1987). Michael Scriven (1987) has also argued that recourse to research-based teacher evaluation involves the "fallacy of statistical surrogation" — relying on indicators of good performance based on correlations instead of criteria of performance based on the definition of the job.

Assessment methods also have come under criticism. The task demands are restricted almost entirely to selecting answers on multiple choice tests and teaching for the record, omitting such other tasks as producing answers in oral or written form, producing materials, or simulating teaching (see Bird [1987] for an appraisal of a sample of current teacher tests). Furthermore, reliance on a few observations of classroom teaching utilizing a standard instrument for counting discrete behaviors is a dubious ground for inferences about teacher competence (see Scriven, 1981; and Stodolosky, 1984).

Efforts are underway in a number of states to develop a more complex approach to teaching quality, based on new assessment methods, elaboration of multiple lines of evidence, and consideration of emerging conceptual frameworks for teaching. Connecticut is experimenting with assessment center methods coupled with tests of subject matter knowledge and with an internship for beginning teachers. Minnesota is considering a three-stage plan to include an entrance licensure test of basic skills and subject matter; a supervised internship conceived as an integral part of teacher education rather than as an assessment opportunity; and a Test of Teaching Skills for continuing licensure that will assess candidates' skills in analyzing and interpreting information about students, selecting appropriate materials, evaluating student progress, and other professional responsibilities (Wise & Darling-Hammond, 1987).

The National Board for Professional Teaching Standards has provided impetus for vanguard efforts. The Board will establish an advanced standard for certification, but in the process will sponsor research and development that contributes a more complex and comprehensive conception of teaching together with new methods of assessment. As the Board's approach to teacher assessment gains credibility, states may respond by adopting board-compatible procedures, drawing on the inventions created by the Board. Already an interstate consortium is in place to convey information to interested states.¹

The effort to create national certification standards for teachers is in its infancy and faces formidable challenges. But for the first time in history, a forum exists for the articulation of advanced standards by the teaching profession itself. State licensure has provided a highly constraining context for the development of standards in teaching. The legal standard defining minimal competence for safe practice, the imperative to supply enough teachers, state traditions of direct legislative involvement in standard-setting, and the prohibitive cost to any individual state of developing sophisticated assessment methods have combined to retard the invention of professional standards and assessment procedures in teaching. If the National Board is successful, then in future the certification standard is likely to exert upward pressures on the licensure standard and to introduce more complex conceptions of teaching into the public discourse.

Concluding Observations

This chapter has proposed a systemic framework for appraising the licensure and certification of teachers. Traditional assessments of state licensure employ legal and psychometric criteria to determine whether requirements are job-related and meet standards of validity and reliability. The perspective supplied here argues that a broader framework is necessary to account for the system of standards that shapes teacher quality in any state. This system must serve the three functions of generating teacher supply, constructing the categories of competence, and inventing conceptions of quality. These functions interact in complex ways and in most professional fields are influenced by the interplay of licensure certification processes, public and professional controls. To appraise the adequacy of any state's licensure system requires attention to these three functions and their interaction.

A framework drawing attention to these functions serves to alert policy makers to the range of factors that determine the quality and quantity of the teacher work force. At present, however, no state is in a position to employ this framework in appraising their credential system. The information base is spotty to nonexistent, making well-grounded judgments impossible. To develop grounds for an appraisal, four steps are required. First, public and professional interests must collaborate in developing a model statute that establishes a norm against which to judge current conditions. Second, states must develop a set of indicators to track progress on the key factors composing the credential system. Third, states must establish a set of procedural standards for regularly appraising the credential system. And fourth, states must build a monitor-and-review capacity to implement the procedural standards. Finally, interstate cooperation in pursuing these steps would be desirable for sharing costs and facilitating cross-state comparisons.

Such an agenda extends the evolution of teaching standards toward increasing state regulation, toward the rationalization of knowledge, the centralization of control, and the specialization of functions in schools. These developments support professional work in public bureaucracies, but provoke misgivings, and so it is worthwhile to contrast this approach with other alternatives.

One alternative is to deregulate entry to teaching, reducing technical requirements and costs of entry in order to expand teaching's recruitment pool. According to this perspective, entry requirements have not demonstrated their utility against the overriding imperative to recruit large numbers of talented, committed individuals from the college labor market. The connection between professional knowledge in teaching and learning outcomes is too tenuous to justify regulation and restriction (Hawley, 1987). Better to allocate resources to inservice education than to preservice requirements, particularly in light of the heavy attrition from teaching in the early years.

Another alternative favors increased differentiation of the profession supported by the licensure system itself. For example, the Holmes Group manifesto

(1986) recommends creation of a three-tier occupational structure: *instructors*, college graduates issued nonrenewable certificates for teaching under the supervision of professionals; *professional teachers* issued a full certificate upon completion of rigorous preparation and examination; and *career professional teachers* issued an advanced certificate upon completion of further specialized study and experience. Entry requirements then might be suitably differentiated, career lines established, and new patterns of interaction among teachers promoted. Within this scheme, entry level standards would vary, and advanced level standards would take shape around the new positions.

The value in establishing a systematic, thorough, and well-coordinated approach to licensure is in creating conditions for the steady advance of professional knowledge and its acquisition by teachers. Proposals to differentiate the teacher corps are also available and the experiments underway eventually may lead to supportive developments in state licensure. Such proposals are compatible with efforts to increase the rigor of the licensure system. Finally, the strategy to deregulate education in the interests of broader recruitment is arguably an alternative worth considering, but some historical evidence indicates that when entry standards were raised during periods of teacher shortages, supply increased (Sedlak & Schlossman, 1986). There may be no causal connection here, but the relationship between entry standards, costs of training, and teacher supply appears to be historically conditioned, not economically determined. The professionalizing trend, then, may enhance supply in the long run.

Note

1. The Interstate Consortium on New Teacher Assessment and Support is funded by the National Governor's Association, with assistance from charter member states Connecticut and California. Currently 32 states have joined the consortium as associates, and this organization serves as a forum to explore emerging trends in teacher assessment.

References

- Ayers, J. B., & Qualls, G. S. (1979). Concurrent and predictive validity of the national teacher examination. *Journal of Educational Research*, 73(2), 86-93.
- Bird, T. (1987, June). *Teacher assessment and professionalization*. Paper prepared for the Carnegie Forum on Education and the Economy. Washington, DC: Carnegie Forum on Education and the Economy.
- Buchmann, M. (1987, March). *Teaching knowledge: The lights that teachers live by*. (Issue Paper 87-1). East Lansing, MI: National Center for Research on Teacher Education.
- Coles, G. (1987). *The learning mystique. A critical look at "learning disabilities"*. New York: Pantheon Books.
- Council for Basic Education (1985). *Making do in the classroom: A report on the misassignment of teachers*. Washington, DC: Council for Basic Education.
- Darling-Hammond, L. (1986). Teaching knowledge: How do we test it? *American Educator*, 46, 18-21.

- Darling-Hammond, L., and Berry, B. (1988, March). *The evolution of teacher policy* (JRE-01). Santa Monica, CA: The RAND Corporation.
- Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. In Ernst Rothkopf (Ed.). *Review of Research in Education*. Vol. 14 (pp. 169-238). Washington DC: American Educational Research Association.
- Hawley, W. D. (1987). The high costs and doubtful efficacy of extended teacher preparation programs: An invitation to more basic reform. *American Journal of Education*, 95, 275-98.
- The Holmes Group, (1986). *Tomorrow's teachers*. East Lansing, MI: Michigan State University.
- Jackson, P. (1986). On knowing how to teach. In *The practice of teaching*, (pp. 1-30). New York: Teachers College Press.
- MacMillan, J. B. & Pendlebury, S. (1985). The Florida performance measurement system: A consideration. *Teachers College Record*, 87(1), 67-78.
- NASDTEC (National Association of State Directors of Teaching Education and Certification). (1988). *Manual on certification and preparation of educational personnel in the United States*. Sacramento, CA: NASDTEC.
- National Education Association. (1987). *Status of the American public school teacher, 1985-86*. Washington, DC: National Educational Association.
- Office of Educational Research and Improvement. (1987). *What's happening in teacher testing. An analysis of state teacher testing practices*. Washington, DC: U. S. Government Printing Office.
- Quirk, T. J., Witten, B., & Weinberg, S. F. (1973). Review of studies of the concurrent and predictive validity of the National Teacher Examination. *Review of Educational Research*, 43(1), 89-113.
- Roth, R. A. (1986). Emergency certificates, misassignment of teachers, and other 'dirty little secrets.' *Phi Delta Kappan*, 67(10), 725-727.
- Roth, R. A. & Lutz, P. B. (1986, November). Alternative certification: Issues and perspectives. Charleston, WV: Appalachia Educational Laboratory.
- Scriven, M. Summative teacher evaluation. (1981). In J. Millman (Ed.). *Handbook of teacher evaluation* (pp. 244-71). Beverly Hills, CA: Sage Publications.
- Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education*, 1(1), 9-24.
- Sedlak, M., & Schlossman, S. (1986). *Who will teach? Historical perspectives on the changing appeal of teaching as a profession* (R-3472-CSTP). Santa Monica, CA: The RAND Corporation.
- Shimberg, B. (1985). Overview of professional and occupational licensing. In J. C. Fortune & Associates (Eds.), *Understanding testing in occupational licensing* (pp. 1-14). San Francisco, CA: Jossey-Bass.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-21.
- Stevens, R. (1971). *American medicine and the public interest*. New Haven: Yale University Press.
- Stodolosky, S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher*, 13(9), 11-22.
- Tyson-Bernstein, H. (1987). The Texas teacher appraisal system: What does it really appraise? *American Educator*, Spring, 26-31.
- Wise, A. E. & Darling-Hammond, L. (with B. Berry and S. P. Klein). (1987, November). *Licensing teachers. Design for a teaching profession* (R-3576-CSTP). Santa Monica, CA: The RAND Corporation.

5. Teacher Selection

Michael Scriven

Introduction

In the broad spectrum of personnel actions, each involving a different set of action alternatives and associated evaluation (and other) procedures, *selection* is the process that begins after the process of *solicitation* has brought in some applications from candidates. Typically, selection decisions are different from decisions about *placement*, *promotion/demotion*, *pay increase/decrease* and *disciplinary action*, but sometimes selection is combined with placement (e.g., when a new school is taking on a large number of teachers and can do some juggling of positions). The selection task is either that of picking the best candidate (a ranking task¹); or of deciding whether a candidate is good enough to be appointed (a grading task²); or a combination of the two.

The Key Questions

1. *What should we look for?* That is, what qualities or achievements are desirable in a teacher filling this position? The answer to this question should come from two or more of the following four sources. Selection procedures that are intended to be serious, fair, and legal require that you determine which of these you will need and develop them well in advance of the recruitment process. The four sources are: (i) *a mission statement* for your institution, which expresses any general defining features such as religious affiliation³; (ii) *A general duties statement for teachers* — either one that is specific to your school, or one that covers most situations⁴ — more dangerously, you may rely on an understanding of the duties of a teacher rather than a statement; (iii) *a job description*, which is specific to this job; and (iv) *a needs assessment*, which is specific to the school at this moment or is at least specific to the teaching needs of this school

AUTHOR'S NOTE: Many thanks to Myrna Tonkinson and Barbara Davis for valuable comments on earlier versions.

at this moment. The needs assessment should be what resulted in the job description, but it is certain to identify some other areas of (service or teaching) need, which have a somewhat lower priority but become important in the later stages of the selection process, as will be discussed further on. Some of this material, or a summary of it, will either have appeared in the advertisement (perhaps as part of the job description or as an introduction to it) or it may be part of a document that is mentioned there as something that any potential applicant should request.

In highly standardized situations, the needs assessment is often superficial: "We just lost our college-track math teacher and we need to find another; dig out the old advertisement." That's almost always a sign that administrators are missing the important management opportunity to redefine the position, even if only slightly, to take account of changed staff configurations and new obligations since the last appointment in that area. But, there are a couple of general issues that need to be looked at, whatever the specifics.

2. *How do we do the looking?* This question means both what general approaches should be used (references, interview, dossier, tests, trials, etc.), and what specific procedures should be employed for getting information about each of the items under the first heading. In particular, it is necessary to be careful not to convert the evaluation of applicants into the evaluation of applications. It is necessary to go beyond the applications, although — contrary to most people's instincts — that does not mean treating the interview as indispensable. But it does include identifying specific qualities in the interview. Otherwise, there is a risk of converting the process into the evaluation of interview performance, a talent much nearer to show business than it is to the classroom.

3. *What do we do with the information when we get it?* How do we combine the various items of information (e.g., by giving different weights to the ratings on the various dimensions)? The combination process must give the answers to two questions: (i) Which is the best candidate (if there is more than one), and, (ii) Is the best (or only) candidate good enough to appoint? Fortunately, the second question is often not at issue. Either the need is so strong that we simply have to take the best candidate (even if only to train him or her); or, in happier circumstances, there are several candidates who are clearly good enough.

4. *Who should do all this?* The leading candidates are principals, assistant or other superintendents, district office specialists, consultants, members of the school board, and other teachers.

Scope of This Chapter

In practical terms, the reverse questions are just as important: What should we *not* look for (of the things which are often sought)? and, What (common) approaches to the gathering and processing of data, and the selection of evaluators, should we avoid? Both the affirmative and the negative kinds of questions must be addressed. In the course of addressing the first set of questions, the

discussion of any one inevitably leads into the later set and into the reverse questions, and we deal with whatever we have to in order to be able to cope with our examples in a reasonably realistic way. Because the later discussions do not repeat all of the points made earlier, they should not be read out of context.

The original version of this chapter covered all the questions listed. The editors' need to preserve reasonable length limits meant that much of the discussion of process (questions 2, 3, and 4) had to be eliminated or severely summarized. It was felt that the topics in question 1 are essential foundations for any later discussion, and they are addressed at greater length. These topics have had less treatment than many others in the teacher-evaluation area, so were difficult to compress. A final warning: this is not an area where the research has been extensive, at least not on the points that are crucial. Contrary to what one might expect, the material on personnel procedures in industry and industrial psychology, despite the vast amount of it, is weak on conceptual matters and on the needed key research. It has, therefore, seemed important to recommend fail-safe strategies in areas where lacking research results or deep analysis.

What to Look for

At the general level, we have to take into account two significantly different — though not absolutely distinct — considerations, and then try to integrate them. The two considerations are: (1) the *merit* of the individual applicants, by the relevant professional standards that apply to the area described in the job description or in the advertisement; and (2) their other sources of *worth* to the institution, that is, considerations that go beyond doing the job as described. Then the results from these evaluations can be combined into an overall rating of (3) their total estimated *value* to the institution, from these two sources.

The relationship between worth and merit is the reverse of the way in which it is usually conceived. The foundation of all personnel evaluation is worth (to the institution), not merit. In the special case where it will be valuable to the institution to have an employee in a certain area, we create and define a job-slot, usually in terms of conventional career paths, because that is how potential applicants classify themselves and so that is how we can reach them through the usual channels of advertisement. We then look for the candidate with the highest ability to fill that slot, that is, the maximum merit. But that is just a convenient simplification. In candidates who will fill the slot equally well, we quite rightly look for other virtues that match some of our needs that do not fit neatly into the standard career definition of, for example, a junior high school math teacher. We may mention some of these in an advertisement or job description, but there is never room for all. The tricky part comes when we ask whether we can legitimately take nonmerit considerations into account to offset a deficit in the merit of one candidate against another. It is here argued that, although there are serious risks about doing it carelessly, the risks are not serious enough to justify failing

to do it at all. We can and should trade worth against merit, within limits that safeguard the concept of open applications. Not incidentally, considerations of worth are what provide the principal, and perhaps the only legitimate, justification for affirmative action.

1. *Merit* means professional merit — the extent to which the individual candidate measures up to the accepted standards of the teaching profession, in the area(s) for which a position is being advertised. The exact nature of those standards is not a trivial matter and will be dealt with in more detail further on. The most important point is that merit here means *future merit*, not past merit. One is selecting for future pay-off, not for a gold-watch award — which recognizes distinguished contributions in the past. Of course, the best *evidence* of merit is achievements — which refer to the past. But one must look carefully for problems about the inference from the past to the future — for example, when someone's past performance, although outstanding, is a long way in the past.

One can always consider long-past performance as evidence of *experience*. That might be useful, as long as it isn't out-of-date experience, which would be a handicap for dealing with new situations. But long past performance cannot be good evidence of probable future *productivity* unless it is coupled with *recent* past performance. One checks in the recent past to see if there is evidence of slowing down to a rate which, when extrapolated, would not be acceptable (or would not make this the best of the candidates, depending on where one is in the selection process)⁵.

Conversely, one must be extremely careful not to weight *promise* very heavily, even when the estimates of it are from very impressive sources. Promise is too often a vehicle for hope on the part of the raters, rather than a legitimate consideration to be weighted against demonstrated performance. What evidence is ever seen about the actual performance of people who were rated as showing great promise? On the other hand, one can immediately think of some people who showed great promise but never fulfilled it; even of people who were appointed on the basis of great promise that never materialized.

Even if there is not much room for promise to materialize — for example, with the older applicant — and even though one is appointing for the future, the mobility figures make it inappropriate to worry about the future beyond five years, which the older applicant may well be able to provide. Promise isn't something; it's only the possibility of something. Never give it any weight in itself. Promise is not a dimension of merit. It is not even an indicator; it is simply an estimate of future performance based on past performance and used when there isn't much of that. Normally, promise is estimated from the quality of the relatively small sample of work available, plus estimates of motivation. It is often better to separate these two dimensions in their own right.

Now, past performance is simply an indicator of what must be assessed, namely future performance. But even if it is not a guarantee of anything, past performance is — if carefully interpreted — the best indicator. Many other indicators are simply illicit.

In some selection situations, there appears to be no alternative but to rely heavily on indicators other than track record in a similar job. For example, it may appear that with a beginning teacher there has not been time to establish a track record. One might then place more weight on their grade record or the opinions of faculty at a teachers college about their "teaching potential" — perhaps based on their sense of the empathic potential of the candidate. Hence, it seems, one not only can use other indicators besides track record, but sometimes one must.

Many of the conventional indicators — an extreme example is the character-reference letter of recommendation — are very much less reliable indicators of future teaching performance than is track record, and some of them should not be used at all. Perhaps character references have a place in establishing trustworthiness with children, especially young children. But even there, character references are extremely unreliable compared to track record. In general, one should almost always find or create a track record before using something else entirely.

While track record is the supreme indicator, it is clear from our earlier example that one must be careful about how the track record is used. Being careful involves particularly: (i) looking at the time-pattern of performance and not just the accumulation of achievements (e.g., looking for overall decline and patchiness); (ii) carefully defining the area in which a future contribution is wanted, so that one looks closely at the relevant part of the track record; (iii) remembering that although there is an immediate need for that defined contribution, it's not the only contribution that one would like, and it may not be the only contribution that one need in the future (in other words, don't forget worth considerations); (iv) remembering that one can't just grab the candidate with the best overall set of track-record qualifications in the superficial sense of degrees and years of experience, because of constraints inherent in the announcement (the terms of the offer) — ethical and legal considerations related to fair employment practice.

Normally the track record is examined for evidence about more than one matter of concern. Conclusions must be drawn about merit in several different respects (classroom management, punctuality, and knowledge of subject matter). It is important to *separate the inferences*.⁶ The literature on judgmental estimation suggests strongly that one will do very much worse looking at the whole record and trying to draw two conclusions from it than looking at the record once for the first conclusion, and then again for the second. Suppose that there is a need for a teacher who will be in charge of computer studies, including the computer lab. On one hand, the need is for someone who will be absolutely reliable in dealing with the security obligations imposed by the supervision of expensive equipment, and on the other hand, there is a need for someone who is keeping really up-to-date in that highly volatile subject matter area. (Of course, there are other matters of interest, such as ability to engage peers in the use of computers, etc., but the point the author is making applies to all further considerations.) The tendency is to read the dossier for all relevant qualities and write down an overall view. It is much better to list the qualities in which one is

interested, read the dossier for each, and rate each separately. While doing this, one is likely to notice other matters that seem relevant. Add them to the list, discuss them with fellow committee members, if any, to see if there is general agreement about their relevance (or at least no overriding difficulties with considering them), and then re-read for performance on these.

2. *Worth* is the value of an employee to the institution, which arises — to be more precise, is expected to arise in the future — from considerations other than their performance of their duties⁷. Hence, *worth* refers to matters not covered in the job description. The two formal sources for standards of worth are the mission statement of the institution or unit and the needs assessment. The mission statement must be assumed to apply to all staff members, as long as they all are aware of it or can be held to be responsible for being aware of it. The mission statement should be attached to any background materials sent out to inquirers, and the existence of (and responsibility for reading) such materials should be mentioned in any advertisements. An obvious example relates to schools with a connection to a religion, but there are many others. Rural schools, for example, benefit from teachers with some commitment to and knowledge of that or similar rural communities. This might figure in a mission statement; but if not, it should be in the needs assessment at most times in the school's history. An exception might be when there are plenty of staff fulfilling that need. A given school always has some specific needs over and above the need for someone in the specific job — a list of them too long to mention in every advertisement and changing too rapidly for any such listing to be accurate. One will recall those remarks in the selection committee's discussions: "It would certainly be handy to have someone who'd like to help out with coaching the girls' softball team"; "A little extra strength in science is really attractive, given our situation"; and so on. These can't form a major part of the selection criteria, because of the need to be fair to those who did not apply just because one did not mention these things. But they can be taken into account when one gets down to ranking a short list.

The line between merit and worth will vary depending on what has been specified in the job description. For example, if the advertisement (or other job description) says "overall contribution to the math department will be given some weight," then we can drop off one of the items listed below.

Thus, examples of worth include the value the institution can expect to derive from:

a. *Performance of services other than the teaching-specific duties.* This might include an obvious talent for and demonstrated interest in working on the school's fund-raising activities. Avoid making this into a reward-for-overload situation by making clear that all teachers have to assist with some of the nonclassroom tasks; here a clear way would be seen in which this aspect of the obligations could be met very well by this candidate.

b. *Benefits arising from incidental characteristics.* Here we have to rethink the usual knee-jerk interpretation of the concepts of affirmative action and equal opportunity. The first woman or minority person applying to be a teacher in the science or math depart-

ments, at a particular school that has no woman/minority staff in either department, may be no better on (professional) merit than the other candidates but may be worth more to the institution because her presence will: (i) make possible types of counseling that were previously unavailable (vocational counseling about handicaps for women in the field, based on personal experience; psychological counseling of certain other kinds); (ii) legitimate a role that too few of the female (or minority) students are electing, by comparison with other, similar, schools⁸. Her presence may also: (iii) be useful to the department in providing representation of the point of view of female students when curriculum considerations come up; (iv) represent the point of view of other female faculty on many issues; (v) have very tangible benefits for the school, for example, in making it eligible for, or more likely to get, certain types of funds or bequests; (vi) increase the social legitimacy of the institution and, hence, its community support, quality of applicants, morale, and so on. While this last consideration is of overwhelming importance in, for example, integrating a police force, it should not be underestimated in legitimating a school or part of a school.

These are all examples of worth, not merit. That is, performance in these areas has nothing to do with her professional accomplishments as a teacher of math or physics, but it bears on her appointability because it bears on her value to the institution. The reason why she scores on the basis of gender is not reverse discrimination or compensatory justice, but a matter of fitting the *overall* needs profile of the school. *It is merely an extension of the fact that a science teacher is going to get preference over a social studies teacher for this job.* The institution needs the science teacher more than the social studies teacher, and that's why they advertised that position. Legitimate considerations of worth led to that preference, and legitimate worth considerations lead to a preference, within applicants that fit the general need (i.e., those who are competent in the needed area) for those who have other needed characteristics.

The knee-jerk response here is in giving extra weight to gender (or, it might be, race), and hence one is automatically guilty of sexism, no better for being reverse sexism. It's true that a simple formula such as "one mustn't give any weight to gender" may be all that most people can grasp, and it's better to grasp that than to proceed without any such guide. At some point, however, policy makers have to take a more serious approach. The correct view is surely that morally opprobrious discrimination occurs when characteristics that are irrelevant to the probable job performance of each individual applicant are taken into account. And job performance for a mathematics teacher — by contrast with a math researcher working at home — does not simply mean competence in doing mathematics, to which it is indeed true that gender has no relevance. In the past, and still today, the most common of these irrelevant characteristics having been religion, race, and gender, normally attention is called to the need for equal opportunity by stressing that those characteristics are not to be given weight. What should be said is that they should not be given weight *unless demonstrably job relevant*; and *job-relevant* must not be identified totally with merit but with institutional needs-meeting. Of course, that will typically require meeting the necessary standard of merit as a first hurdle; there is no justification for letting *worth* considerations replace *merit* below the level that is required to get the job done.

It is not enough to say that only when two candidates are *equal* can we appeal to worth considerations. If two candidates are well above the threshold level of merit to get the job done well, then we should look at their total contribution to value, including their worth as well as their merit. It is simplistic to suppose that the value of still further merit

points, and hence of differences in merit, does not begin to decrease after a certain point. As the marginal utility of merit diminishes, the relative importance of worth goes up.⁹

Thus, the usual cry of "appointments must be made on the basis of merit," typically seen as a cry for justice, is calling for something that is only superficially equivalent to justice. If appointments were made on merit alone, one could never hire a math teacher if a more meritorious music teacher applied. Appointments must be made — and staff retained — on the basis of worth, not just merit. One appoints the math teacher because one needs a math teacher more; one appoints the female or minority person because one needs them more.

Now, one can't play fast and loose with this. It does not automatically mean that religious schools are entitled to require religious commitment from their staff, any more than it means that girls schools are entitled to restrict their appointments to women. What it does mean is that such restrictions must be treated as perfectly possible in principle, not as violations of some fundamental principle of justice. In a particular case, one must look at the whole needs assessment picture, the market situation, and the worth of diversity as well as of consistency.

Legitimizing considerations of worth is seen by many as letting the camel get its nose under the flap of the tent. For example, suppose that a wealthy member of the community gets into a sustained argument with a probationary teacher, who is a leftist political activist, and offers the school a swimming pool which they sorely need, if they will refrain from giving this teacher tenure when that issue comes up at the end of the year. We are all repelled by the idea of giving in to this kind of pressure, but doesn't the legitimizing of worthy considerations force us to do so?

After all, as the donor points out, the market being what (one can hypothesize) it is, one can easily replace the teacher with someone of equal or greater merit, and the pool will not only have exercise benefits for the whole school community, but will support a swimming team which — so the sports director and several delegations assure one — will dramatically enhance school spirit. Once more, one has to look more deeply than usual at the situation; the answer does not lie in saying that nobody should lose their job unless they fail to perform their duties satisfactorily, the usual union slogan. It's clear that people should lose their job when the need for the job evaporates — although not overnight. Schools, no more than companies or governments, cannot be held responsible for foreseeing all demographic or market changes, and taxpayers should not have to support noncontributing staff beyond a reasonable relocation and/or retraining period.

The line of argument that one should stand by, at the cost of a swimming pool, can be put in one way as the consideration of precedent, which is so serious in this case that it could cost many of your best teachers. Many will not stay in a school where their citizenly activities can lead them to being sold out (literally)¹⁰.

The difficult cases are not restricted to extreme ones such as that described. Does the notion of "upsetting the community" ever enter into selection decisions? Who has not heard some mention of that ominous phrase "troublemaker" in personnel discussions? Isn't this a worth consideration? The euphemisms of "lack of support" or "serious concerns on the part of some members of the school board" will often decide a case for a timorous superintendent. Of course, there are extreme cases to which these remarks might apply, for instance merit or noncontentious areas of worth; but not always. Based on discussions of personnel procedures over the last few years, there is a good deal of

evidence for the view that discrimination on the basis of age or nonconformity is now more common and less recognized than discrimination against minorities and women. However, the existence of difficult cases is no justification for retreating into oversimplification. The weighting of worth cannot be dismissed in favor of a pure merit approach, even in the type of case discussed so far, and one or two other cases should be considered.

c. *Ability to perform services that may be needed in the future.* This includes the value of someone with extra qualifications that would make them useful as reserve or back-up in another area. The limiting case of this type is the teacher who is fully certified and experienced in two areas, and who may thus immunize one from the problems and costs of redundancy proceedings (or inappropriate placement), should their primary area (in which one is now considering their application) cease to attract enough students to support the number of tenured staff one will have on-site.

d. *Interactions.* When the needs assessment was done for this job announcement, undoubtedly, the gaps in the present staffing profile were considered. One may also have looked at, and perhaps made some mention of areas where there are no gaps in the obvious sense, but where there is a specialization that needs further strengthening in order to achieve critical mass. However, the line was probably drawn at mentioning that it would also be an attraction to have a math teacher who could serve as an assistant coach in basketball, as there is no coach in several sports for which there is some demand. It is simply not practical to list *all* the gaps and *all* the places where further strength would be valuable, let alone their relative importance, in all possible strengths and combinations. (Note that *interactions* does not refer to *probable harmony in the staff room*; that's not what teachers are hired for, and it should be given low weighting against good teaching ability.)

Hence, there is always some need to do pattern matching when the more promising candidates' specific abilities are known. We often talk then about "a particularly good fit to our needs" or, on the other hand, about a candidate who "doesn't seem to contribute much to the department — apart, of course, from their ability to fill the specific job requirements." This is another example where the minority or female (or, in some cases, male) candidates may have virtues because of their membership in those groups, in bringing another point of view to bear. And it is another example where one walks on thin ice, moving into the area where the community begins to have thought-police control over school teachers; more on that in a moment. Notice that *interactions* does not refer to social relations between members of the staff; that should get a very low rating — it's essentially a bonus for those already appointed, not a benefit for students and taxpayers. It is discussed later.

e. *Community relations.* Something has already been said about this, but it deserves to be set out explicitly. To take a clear case, being well known and liked in a district because one grew up in it might make a candidate very valuable to a school that needs help in that dimension. If possible, it should have been mentioned as an advantage in the announcement of the job; but one usually can't mention all relevant sources of worth.

Distinguishing Merit from Worth

A simple case that sometimes proves confusing is the over-producer who does much more than the duties require. This kind of probable over-performance of

duties, in the primary quality or quantity dimensions, as in "He's worth two ordinary workers," is simply a case of merit. Merit always has the two dimensions *quality* and *quantity of work*, and this is a case where the quantity score is very high. It should be picked up in the discussions of merit, rather than under the heading of worth.

In the light of these arguments, which appear to establish the ultimacy of worth, an obvious question arises, why bother with merit at all? The answer is threefold. First, there is an effectiveness gain in getting applicants. More and better applicants are reached if we target our announcements and commit to a merit-focused process, which we must then follow in the selection phase. Second, we can operate the selection process itself more effectively if we use a two-stage approach, just as we always begin the selection-of-consumer-products with developing a short list of the best options (on plain merit), and then look more closely at these. Third, there is a deeper ethical reason, analogous to the answer to the vigilante's question: If the justification for the law is that it benefits the society, why shouldn't we act contrary to the law when it is obviously better for the society (e.g., by killing a killer who is going to get off in the courts because of some trivial failing to follow procedure)? One way to put the answer is that staying with the processes of justice - even though they sometimes cost us something - provides the best insurance for the preservation of justice. It's simply not true that really smart people don't pay insurance premiums. Just as the vigilante wants to short-circuit the safeguards of justice, the idea of simply selecting on overall worth would lead to very high unreliability in the process, and, hence, to injustice as well as inefficiency.

However, it is true that smart people don't pay annual insurance premiums that approximate the cost of the insured property. And it would not make sense to say that worth should be rated as much than merit, *after* the determination (based on worth) that this particular position is the one that most needs filling. That determination gives worth most of its due and is covered as soon as a candidate meets the minimum specifications for the job; beyond that, we should rely more on merit! That is, the limit on the weighting of worth as between qualified candidates is surely that worth cannot outweigh merit, and would rarely come anywhere close.

Balancing Worth Against Merit

This is of course an issue from our third category (What do we do with the data when we have it?). But it relates to a single overarching issue that we need to get out of the way before going on to talk about issues that are more teaching-specific.

Without abandoning justice, how much can one weight worth against merit? To summarize, we have identified four positions which it may be helpful to label.

(i) *The No Trade* position (only merit should count), whose serious problems we have already spelled out. (ii) *The Other Things Being Equal* position (only if

there's a dead heat to merit), which the preceding arguments suggest is still too limited. (iii) *The Reasonable Balance* position (trade-off is legitimate up to the point where the cost of and possible losses from re-advertising are less than the probable gains, but never past the point of weighting worth as much as merit¹¹), for which we have argued; and (iv) *The Free Trade* position (merit is just one way to get value, so all trading is legitimate), which runs up against the argument just given for a two-stage process. The Case of the Swimming Pool Bribe is also a counter-example to the Free Traders. It shows that the mere size of the worth gains cannot override all considerations of merit. Vigilantism is a threat to justice, partly because of the precedent it creates; it is not, contrary to what seems obvious to its advocates, just a highly practical implementation of justice.

The Reasonable Balance position, as supported above, is intensely practical, too, but not quite so radical as the Free Traders. We've formulated it in terms of allowing trade-offs when: (i) both candidates are well above the level of merit needed for the job; (ii) the weaker one on merit will probably contribute more in terms of worth than the marginal benefits from the difference in merit; (iii) the gains from re-advertising — ethical and practical — are less than the costs, in lost time, morale, and availability.

The Reasonable Balance position would be made more effective if graduates from teachers colleges were given some understanding of the whole process of how jobs are advertised and candidates ranked. They can then see that one should apply for jobs that are close to their area, if they have other strengths that will count as worth and could be mentioned. In general, the situation should be improved by improving the extent to which potential job seekers are taught this, rather than by requiring something which is probably not possible in principle, certainly involves major delays and wastage of resources (panel time, cost of advertising), and may lose some of the best candidates, although it may also pick some up.

It will also help achieve efficiency and justice if we can include an account of the more obvious environmental variables — the ones that bear on worth — in the materials sent out to applicants. This sometimes involves a list of other members of staff and their interests, special concerns of the school, etc., coupled with a general remark about the likelihood that some weight will be given to *fit*.

The Qualities We Should Look for in Teachers

We should look for five qualities in particular, of which the first two are rarely given the emphasis they deserve. There's nothing startling about their headings; but the details are not being implemented, partly because the market is not overflowing, but to a much larger extent than should be the case because they are not being taken seriously by selection committees¹². It is naive to assume that credentials from any teacher-training institution afford a guarantee of competence in even one of those of five areas. The basis for this list of qualities is the

list of legitimate sources mentioned at the beginning of this chapter, especially the job description and the generally understood requirements of the teacher's job — the duties list. These qualities are merely listed now, but we will refer to them when discussing procedures such as interviewing.

A. *Sound knowledge of subject matter*, including the across-the-curriculum subjects (literacy, social/personal/study/work skills, computer competence); so that what is taught and learned will be correct. For the secondary teacher this makes a degree from a reputable institution *with a major in that subject matter* almost essential; degrees in "math ed" are, in general, completely unsatisfactory substitutes. It's not that some of the best math teachers in the world have any more than that, it's that in the selection process one can't depend on any less in the form of paper credentials.

B. *Solid competence with and understanding of testing* and/or structured observation; so that the teacher will find out what learning is needed early in the term, whether their teaching is working, and provide timely and accurate feedback to students, staff and administration about students' progress. Sample tests authored by (not just selected by) the candidate should be provided and discussed.

C. *Classroom teaching ability*, including classroom management skills¹³, A-V and other educational technology operating skills, and teaching materials production (this applies even for beginning teachers). This should be demonstrated in person and on site if it cannot be established beyond question by at least two independent and experienced observers.

D. *Teaching-related intellectual and personal qualities* (professionalism, a lively intelligent mind, punctuality, conscientiousness, willingness to help other teachers, and commitment to teaching). Note that *apparent enthusiasm for teaching* is not allowable, although enthusiasm is in general a characteristic of good teachers; it's simply a style variable.

E. *Worth*; the value to the school that comes from courses other than professional competence (this will vary from situation to situation, but often includes such factors as versatility, familiarity with the type of community, etc.)

Matters That We Should Not Take into Account

These are qualities about which we should try to avoid getting information, and if we come by such information, we should avoid giving any weight to information about these matters. This means actively avoiding the acquisition of this material, and it means accepting responsibility for preventing any such information coming to our attention, where it is possible to avoid it. For example, the chair of the interview panel must cut off any line of discussion that is heading in the taboo directions; letters of recommendation that mention any of them should be copied and the copies censored by someone not on this committee, and so on. This kind of information is lethal, because if any of it comes to light, it may affect the judgment of (some members of) the committee, and it can't easily be proved not to have had this effect. Hence, it may be argued that it contami-

nated the whole selection process, which would make the decisions appealable. It is impossible to exclude all of this information if an interview occurs, which is a serious reason for excluding interviews. But if it does turn up, as some of it inevitably does in an interview (e.g., age, sex appeal, or job-irrelevant personality quirks), then it should be at least be zero-weighted in an *explicit* listing of considerations that is carefully followed as the basis for the selection decision. The following list is normative only; there will be exceptional circumstances in which these matters could be shown to be job-relevant.

A. The candidate's family and personal life or its absence. (When the personal life shows up in work deterioration/improvement, of course we take note of the latter; but the causes of work change are none of our business. Nor should the candidate be encouraged to offer explanations in terms of personal lifestyle changes, nor should any that do come up be given significant weight, as one is not in a position to check on the accuracy of the report nor on its probable effects with this individual.)

B. Any particular type of personality or appearance (gregarious, leadership, charming, fashionable) short of those that are demonstrably essential for, or incompatible with, performance of duties. In general, it is a fault in an employer or coworkers to be unable to use someone because they do not find their personality attractive. (The employer is supposed to be maximizing performance of duty, not pleasure or easy social interaction with coworkers; and abrasiveness, for example, may lead to better results than friendliness in terms of improving the quality of output.)

C. The candidate's sex, race, religion, or age, except in the rare cases where it is justified by the discussions of worth, for example as above.

D. Any written or spoken performance on topics where it is quite clear what the desired type of answer is and there is no independent way of verifying the truth of the answer given. A common example is asking someone to "tell us about your reasons for wanting to teach." Because confidence artists can do this at least as convincingly as sincere people, and interviewers are not expert at discriminating between the two, it is an invalid basis for decisions.

E. Any particular teaching style (e.g., positive reinforcement, inquiry approach, use of detailed lesson plans). Teachers have the professional right to select any teaching style that does not render the duties of teaching impossible and to be judged on their success with this style, not on the evaluators' style preferences. This comment needs immediate justification, as the opposite practice is very widespread. Teachers are often judged by the extent to which they conform to what might be called the *researcher's model* of a good teacher; sometimes by their conformity to the principal's model. Now in a private school, if the trustees do not object, this may sometimes be justified. It is not appropriate, indeed not ethical, in state-supported schools, for the following reason.

Research on teaching allegedly shows that certain styles tend to be more successful than others; what we are calling the *researcher's model* simply combines all the allegedly winning styles. This approach involves the same fallacy as the judgment that because the crime rate is 6 times higher among redheads than brunettes, a brown-haired candidate should be favored over a redheaded one for

an administrative position. You must judge candidates by their own achievements, not by the achievements of those who resemble them in some way¹⁴. The problem with this kind of inference has nothing to do with the unsoundness of the underlying generalization (it may well be true), nor is it simply a moral or legal point. It is simply that using the best evidence we can get, which is track record evidence, immediately invalidates the application of the generalization. If we know Luis Mendez is one of the top three money makers on the pro golf circuit, we can't suddenly suspect he's not a good player when we first see him on television and discover that he has a rather clumsy looking swing. Even though we are quite right to believe that people with that kind of swing tend not to be good players, our generalization can only be legitimately applied when we lack more direct evidence of merit.

How to Get the Information We Need and How to Use It

We go about the selection process by using a selection from four data-gathering processes, in addition to (i) the application from the candidate containing the usual details. We may: (ii) request submission in writing of further information/documentation by the candidate; (iii) seek further information from referees, former employers, supervisors, colleagues, students, or teachers; (iv) interview the candidate in person or on the phone or by proxy; and (v) arrange tests or trials. It's mostly obvious how we get most relevant items of information from one or another of these sources. We'll focus on the less obvious aspects of the process. Remember that following a careful procedure like this, and keeping the notes which show that we did so, establishes an "audit trail," which is our best defense in any case of appeal.

We begin with a general caution. All key data should be independently checked. (Key data are material without which the candidate would not get the job.) This is not the large work load that it sounds at first, as it is only relevant for candidates that get onto your short-list. One may not yet have run into a case of forged credentials, but any professional personnel officer knows that they are all too frequent. Their cost to a school can be very high; and the cost to the students can be appalling, as some well-publicized recent cases have made clear. This precaution means that one must make some calls other than to the named referees. A secretary can make the calls to the registrars at institutions from which academic credentials are supplied; but they must be made. It is not only important to avoid the embarrassment of the situation where it turns out later that the credentials were distorted or forged, but potentially dangerous to the students. Recent examples of university presidents who turned out not to have received the doctorates they claimed come to mind; those cases get some attention, but the rest of the iceberg is there.

Next, as already mentioned, it may be necessary to exclude some of the information that often finds its way into the standard dossier — and disregard

other information that we cannot exclude. A trained assistant should be used to selectively remove this information, leaving a trace so that an explanation of why it was removed can be given. (This is an uncommon procedure, but it is sure to save one trouble in an appeal, and can save the case.)

Given that we now have purged dossiers, it is possible to hold a first meeting (in the case of a single evaluator, this just meaning doing a first serious reading). This meeting should aim to identify three groups: the *no-way* group (one might use the mark X against their names on the list, and refer to them as *crosses*, for *crossing off* the list), the *maybe* group (marked with a ?, and referred to as *possibles*), and the *very promising group* (*, for *stars*). In addition to the first mark against a name, there may or may not be a second mark; it's convenient to have this in a second column. If there is, it is always a question mark, and what it means is that some key data are missing or need to be verified by some means short of interview (i.e., by methods (i) to (iii) on the list that begins this section).

At this first meeting, all stars get a question mark after the star, because it is essential that the key data on them be verified. (This first list of stars is our first cut at a short list, but it's very tentative.) The maybe group consists of candidates of two types; for some, all the needed data are in, and the conclusion is that they are marginally possible appointees, but not preferable. They get a single question mark. Only if stars do not survive the verification check will we do one on the possibles. The second sub-set of possibles still lacks some data, although your tentative judgment is that they are no better than possibilities — and no worse. They get two question marks, and are not finally classifiable for merit until some further data is obtained. Similarly, some of your weakest candidates have a lone cross against their name, but others have a cross and a query meaning that there needs to be some further check before they are finally rejected. (Theoretically, there could be a category of applicants about which we can make no preliminary assessment; these would have a blank in the first column and a query in the second. This is extremely rare in practice.)

At this meeting or reading we need to identify the missing information that must be found in order to remove each of the question marks in the second column; put a note about it on the cover of the relevant dossier. Even if there is a long list of (tentative) stars, one shouldn't skip the follow-up on the possibles, unless it is clear that the best possible outcome of the further inquiry could not make them competitive with several of the stars (several, because not everyone offered the job will take it).

The second meeting is held the moment that all queries in the second column have been removed. (If we have a long list of stars and possibles, we might not worry about removing the queries on the apparent no-way candidates¹⁵. We will now be able to see how to dispose of the possibles (they get a cross or a star, or remain in the doubtful category), and whether some of the original stars should be dropped out or dropped back to possible status because their credentials did not check out. A solid short list should come out of the second meeting. The short list is usually defined as the set of candidates that look as if they (or at least the

best of them) should be further investigated by using one or both of the two remaining methods: interview or special tests¹⁶.

These first two meetings can sometimes be compressed into one by an experienced administrator, by using the phone — if there is no time-zone problem. (Hence, the meeting should be carefully scheduled if an attempt to do it all in one meeting is to be made — or some calling could be done by the committee chair in advance. That will involve some unnecessary calls as the committee is not likely to see things exactly as the administrator does.)

We have already listed the types of information that we must have, from the dossier or from other sources. It's time to say where it's going to come from.

1. We must have information on classroom track record, because anything else is such a poor second (e.g., scores on national teacher tests, grades in teachers college), and if we find ourselves unable to get it from the obvious sources (supervisors in previous schools), we should set up trial teaching periods in the school district in order to get it. However, it is almost always possible to get it; in the case of a beginning teacher, we can get it from their supervising teachers wherever they did their practice teaching and from their supervisor in the school of education who presumably visited their classroom. Special trials are not ideal, because the teacher is going in cold, is not there for long enough for observers to pick up a sense of their conscientiousness and perseverance, and is under atypical stress. But trials are much better than no data or data of really unknown validity (e.g., recommendations from someone we can't reach for further discussion); and they are asymmetrically useful (that is, someone who does well in them should be regarded as having, at least, high peak capability, even though someone who does not do very well may only fail due to nervousness in an atypical situation). The asymmetry is presumably no worse than that for an interview and relates to a different and extremely important area. Furthermore, trials have the advantages of relating to the school where the work will be taking place; and they can be observed by the people whose standards have to be met. But there is currently a serious problem with classroom visits — the wrong kind of observations are made, especially observations of style rather than just of success and content.

It should be mentioned that student ratings, if gathered in a suitably secure way, using a suitably designed form, can provide a useful basis for rating teachers. They are extensively used for this purpose at the tertiary level, and some tertiary institutions have been so using them for 40 years. Their use is still rare at the secondary, let alone at primary, level but they should work there if students and teachers are prepared for their use. In general, they are worth a great deal more than the usual reports from occasional visitors¹⁷, whose reports suffer from samples that are inadequate in size and not representative, measurement artifacts, style bias, and failures of empathy, and are usually vulnerable to personal bias.

2. We must have information on subject matter expertise, and the information must show that the subject matter expertise is up to date. That means it is not enough to have evidence of college courses a long time ago. As a rule of thumb, these qualifications become senile after three years (in computers) and perhaps not until ten years later (in foreign languages). That means you really need information on continuing professional development in many cases. This might be evidenced by further courses, documented reading programs, demonstrably active study group or user group membership (the

evidence of activity might be papers presented), and recent class materials. This information may be available on the application, or by request from the candidate (or previous principal or head of department or — weakest — from previous colleagues or instructors at continuing education courses). In any case, such knowledge and material would have to be checked over by a subject matter specialist and discussed with the candidate to ensure authorship. To avoid spending time on this in the interview — or adding stray characters to the interview panel — one may want to arrange a pre-interview with such a person, the results from which will come in to the interview in a form that can be quickly assimilated. This is one area where it is best to avoid reliance on one referee; two is a better minimum.

3. Exactly the same applies to testing skills and knowledge, here taken to include all structured observation approaches to assessment, and knowledge of and good practice in methods of reporting test results. There should be comments on this in the letters of recommendation; if not, one needs to call the referees along with previous principals, for track record on this ability. Next best; have a specialist on staff do an interview at which examples are discussed rather thoroughly with the candidate. This has to be one of the areas of greatest weakness in the professional arsenal of teachers today.

4. We must have information on the out-of-class contribution to teaching, especially committee work, level of effort, aid to the profession, and counseling. We can get all that in the same way as the preceding. Notice that so far we have not mentioned the interview as anything but a last resort.

5. Finally we come to the matters for which an interview is most helpful. First, it has an important verification function. It can verify that competence or excellence in the preceding matters is in fact in the applicant's repertoire and not just in his or her helpers' repertoires. Second, it has a direct test function in connection with a number of skills. Of these, perhaps the most important are: (i) language mastery in the mother tongue (and any others that are relevant to merit or worth); (ii) the ability to explain matters clearly (at least to peers) and (iii) to cope with questions on one's feet (and in a high pressure situation, which is especially job relevant if any future as an administrator is envisaged); (iv) familiarity with important current issues that affect the teaching profession and the schools; (v) knowing something about the particular school (or at least the district) to which they are applying — or at least showing interest in acquiring that knowledge immediately; (vi) exhibiting teaching-related intellectual and social qualities such as enthusiasm about their subject and about getting students interested in it¹⁸, having some idea of their own strengths and weaknesses; having some ideas about their own future, especially including plans for continued professional development. Conversely, and of almost equal importance, it should be an opportunity for the candidate to discover the answers to some questions about the intellectual quality and values of future colleagues, and about school and community resources.

Problems with the interview

The interview can be useful, if well managed, but it is important to get an early warning into our discussion of it: interviews are generally more of a problem than a solution. The interview is the Rorschach test of the selection process, something onto which different interviewers project their own interpre-

tations. Yet it is highly valued, especially by those who are not familiar with the literature on its unreliability and invalidity. Its supporters are not wrong to think that valuable information can be garnered from an interview, as the preceding list shows. The problems are; (i) to avoid mining in the wrong places; (ii) to distinguish the gold from the dross in what you dig up; and (iii) to avoid contamination of the gold that you do find by the poison that should not have been turned up. It is certainly not true that most interviewers' instincts are a good guide in those difficult task. Often the very conclusions about which they feel most confident about are the least accurate, as we have found in the course of back-tracking students or staff that have turned out badly¹⁹. And some of the data that interviewers think really enlightening are rightly identified by the affirmative action officer as improper considerations. Given all this, one has to be skeptical about almost all evaluations that are heavily dependent on interviews.

It does not help matters that there are a number of hidden agendas in many interviews. For example, interviews are often used as an opportunity to show off one's own skills in front of colleagues — perhaps one's subject-matter mastery, or one's outstanding skill in the chair, one's world-class bonhomie, or one's superbly humane treatment of the allegedly hapless interviewee. And most of us will have seen interviews where the questions to the candidate were used as devices to further a long-standing feud between cliques on the panel.

Interviews are, one should remember, the chosen battleground of used-car salesmen, when what we need is a warranty. Interviews are the province of the peak performer, when what we need is a stayer. Nobody shines in an interview better than the psychopath, and the usual interviewers for school jobs are surely not competent at identifying psychopaths in an interview. It's well known in the clinical field that it's extremely difficult for highly trained clinicians to identify psychopaths from single interviews, without test data or long-term observations by a clinically trained professional, with reality checks.

The two main problems with the con-artist are: (i) the impression that is made by their charm and fluency; and (ii) their ability to subtly manipulate the direction of the interview. We can do something to control these risks, although we should not suppose that doing so lifts the interview to a high level of validity.

In order to reduce both the charm and the manipulation effects, even with face-to-face interviewing, it's important to prevent the interview turning into a free-ranging conversation (attractive though that alternative is to many panelists). Conversations run us into three kinds of trouble: the two just mentioned, and they also become highly noncomparable between candidates, which creates acute problems about equal opportunity. (How would you react to the idea that we should give all candidates written tests that were substantially different among candidates?).

Clearly, the safest procedure involves having the chair of the interview committee run it with a firm hand. The chair may start with one or two standard questions, then turn over to others the task of asking some preassigned questions and carefully chosen follow-ups on them. The chair must stand ready to head off

either interviewers or candidate if there are signs of straying into exotic territory. This strategy is greatly assisted by using a predetermined and prewritten list of questions that must be covered; by avoiding the idea of a fixed time limit on the interview; by trying to arrange for a few minutes of recess — during which the candidate withdraws — while the committee reviews progress and possibly refocuses; and by trying to ensure that there is a possibility of a second interview if necessary.

Exercising the necessary tight control over an interview is particularly difficult if the panel is large or contains people senior in rank to the chair. In the personnel business, the general view is that a panel of five has reached the danger level. Three or four is probably ideal, as long as they are not dominated by one person. If they are all to have their say about new points that come up in the interview, it becomes very difficult even with a small panel to ensure a reasonable degree of uniformity. But one can let each of them ask their favorite question of each candidate and keep score of comparative performance on those items. In terms of numbers of candidates to keep track of, again the process is much easier and more reliable if there are two to five on the short-list; although sometimes a larger number is unavoidable.

Another problem with an interview, according to equal opportunity officers, is that women and minorities have to tread a much narrower line in terms of acceptable behavior than do white males. They have to take extra care not to be seen as too uppity on the one hand or too inert on the other. The generalization of this point connects to much we have argued above: that interviews place altogether too much emphasis on personality style and not enough on work achievement.

One way to put the problem is in terms of a confusion of contexts. Meeting people and appraising them while we do so is something we do socially and extensively. In that context we are interested in whether we like them or find them attractive. It's all too easy for the average panelist, whose principal background is social "interviewing," to treat evaluating candidates in an interview as if it were the same kind of task. Indeed, listening to a panel talking about how important the interview will be in making a choice often sounds as if one is listening to people trying to convert personnel evaluation into something with which they are familiar. Unfortunately, the kind of social appraisal with which they are familiar is not aimed at, and is not verified in terms of, its success in identifying good workers. This lust to interview is illicit.

But, it may be objected, aren't most interviews done by future colleagues, and isn't it appropriate that the interview be aimed at identifying candidates that will make good colleagues? That is, surely the social evaluation is relevant? Out at the margins, this seems obviously true; someone who is so aggressive that they won't allow the interviewers to get a word in may be beyond the pale. But out at the margins we nearly always pick this up before interviewing, with a couple of serious phone calls. Apart from such unlikely cases, we have to decide whether we are selecting for likability or productivity, and although not a complete

dichotomy, there's more difference between the two than most people realize. Take a good look carefully at the best workers currently on staff. Often — and we only need one or two cases to make the point — these are not people who are highly gregarious or even viewed as attractive or as good friends by many others. Nor is this different for people who are not good workers. The difference, if any, is likely to be in the respect of others, not their liking. So the rule is: interview only for qualities that you can respect, not qualities that you like. Some people can do this; others find it difficult to avoid the cult of personality — as is quickly apparent in their comments on what they learned from the interview. They are looking for possible friends or for someone attractive, not for successful teachers. It will be seen from this discussion that the interview is a risky business. It is worth considering some radical alternatives.

Telephone Interviews

The visual part of the charm effect can be reduced by using phone interviews, and using them also has a large benefit for the budget (making interviews possible in cases where they would not otherwise be possible, as well as improving their purity when they are conducted). It can also result in big contractions in the decision time-line (everyone can be reached for a phone interview within a day or so, whereas it may take a week or more to import other-coast or overseas candidates for in-person interviews). For some candidates, it reduces anxiety; for all, it provides a better test of the clarity of speech in English than does a face-to-face interview.

One valid objection to a phone interview is the fact that the candidate does not have a chance to see the site (if they are not already familiar with it). But that can be done after the offer is made and before it is accepted, if it is important to the candidate; and details about labs, etc., can be provided in writing and by using photographs or an interview with someone already using them. Another common objection is that one doesn't get the same sense of presence that working colleagues and students will have, so the phone interview is unrealistic in that respect. But what is it about presence that can legitimately influence you? Not appearance. In the rare cases that that is beyond the pale, in some respect that is job-relevant, the referees will usually comment, and you can make sure they do by putting it on the list of questions to ask them.

None of these suggested constraints need eliminate the opportunity for the candidate to talk about his or her interests in an informal way, or to ask questions about the job or site that are of concern. Nor does the structure suggested mean that interviewers can't follow up on a lead, for example, look for evidence of self-development or outside strengths. And all of this can be done in a phone interview. Here are one or two other suggestions about the phone interview.

- (i) Using a speakerphone is preferable to using a conference call, because it allows for visual and written communication between panelists during the interview²⁰; in fact, such communication is more practical than in the presence of the interviewee. Signals can be used to inform the chair that more time is needed or that the issue seems resolved favorably (or unfavorably)²¹.
- (ii) It's highly desirable to have someone, preferably the chair, do a full try out of the equipment in advance.
- (iii) The chair should open by introducing those present, to give the interviewee more of a sense of dealing with real people, and encourage the interviewee to write down the names and their location around the table.
- (iv) Allow at least ten minutes more than the time you think you need for the interview, to cover reconnection times and other mechanical troubles, as well as quick discussions of impressions while they are still in short-term memory. Unless very limited matters are at stake, 30 minutes seems to be a good average time to allow for interviews; some will run more and others less. If the signals suggest that a candidate is close to being dropped off the list or to topping it, be particularly careful to ensure that all relevant points have been covered, and any new ones that have been added to the agenda in earlier interviews with other candidates, in order to avoid inequity.
- (v) Tell the interviewee to call the chair back with any afterthoughts, perhaps at a specified time, or perhaps across a period of several days. Tell the committee that they, too, can request a reinterview at those, or — if necessary — other times, if they have afterthoughts that persuade the chair or a group of committee members of their importance.
- (vi) Before the interview ends, be sure to ask yourself whether there has been a chance for the interviewee to show depth of thinking as well as good speed of response to standard questions; possible questions are — What pedagogical problem have you been facing/thinking about lately? or, What topic or subject-matter problem have you been working on lately?
- (vii) Recording interviews, whether on the telephone or not, is relatively unexplored territory, but might be considered on an experimental basis, if all candidates agree. There is a significant risk of inducing unease, but it would reduce disputes about what was said; ask yourself how often this has been a problem in your experience.

Is there a problem of equity if we phone-interview some candidates and personally interview others? That will depend on the care with which the interviews are done and interpreted; it is not a necessary consequence of the situation. Even if studies, not yet done, show that there is some statistically significant tendency in one direction or the other, that would not show that bias was built into the procedure. There are many other factors that could explain a difference (e.g., nearer candidates are likely to be more familiar with the job situation, which may be a valid reason why they turn out to be more likely to succeed; further away — the ones most likely to be phone-interviewed — means more likely to be engaged in postgraduate work, hence better educational qualifica-

tions). Committees should try not to be prejudiced by the difference between the interview modes, and unless they strongly disavow their ability to operate in an unprejudiced way, one might try the procedure out. It is true that, in the light of what has been said above, there is a strong possibility of differential treatment, but this might be offset if it was clear that the best phone-interviewed candidate would be brought to the site for an interview before a job was filled. In that way, we control for the possibilities that those interviewed by phone either receive a substantial advantage or a disadvantage from the particular panel we are using.

The most important suggestion is that we should avoid giving substantial weight to general impressions from an interview that are not specifically related to the list provided above. Analogously, we should avoid giving strong weight to judgments by referees, however experienced, who may be reaching beyond the factors listed, for example, who may be reacting to teaching style.

The Use of Trials and Tests

A trial run at teaching a class is a very severe test; many teachers talk of taking six weeks to get a class into a sound working relationship, so one day . . . but of course, the substitute teachers do it every day, and some of them do it very well. It's a test that can give you a winner, but shouldn't count too much against someone who looks good in other respects, as long as the job is not tenured.

An equally serious matter is the literacy test. It really is important to make sure that literacy is verified, because many teachers whose letters of recommendation and credentials are impeccable are in fact a source of embarrassment because of the spelling and grammatical errors in the letters they write to parents, because they can't pick up errors in student work, and because they often put misspelled words on the blackboard. By asking the candidates to demonstrate their proof-reading ability to a sample of student work, and to illustrate how they use the blackboard/whiteboard and overhead one can go right to the most important of these skills (for the job of a teacher) and incidentally give a very good spelling test. Only you can decide whether the possible awkwardness of this practice — because it is rarely done — is too high a price to pay for avoiding later embarrassment.

How to Combine the Information

The usual teacher selection process — like many other personnel selection processes — conventionally and sensibly involves dividing the task into two stages; the first is the identification of a short-list and the second is the identification of the eventual winner(s) from the short-list. This two-stage approach makes good sense as long as one doesn't cast it in concrete as a required approach. It should not be used if there are real signs of difficulty in getting agreement on dropping anyone off the short-list. And one should use a three-

stage approach when there is a huge number of applicants. In most cases, however, the two-stage approach has the advantage that one can use much quicker evaluation techniques for the first cut, as long as they are fail-safe, that is, they avoid false negatives (i.e., rejecting someone who would have been good). Of course, any procedure that avoids false negatives increases the chance of some false positives, but those errors are less costly since we'll pick them up when we get out the magnifying glass for scrutinizing the short-listed candidates.

The most important way to make the cut to the short-list is to look for the absence of absolute necessities, such as poor subject matter or classroom control competencies²². It's helpful if the evaluation panel can spend a little time identifying these absolute requirements at the very beginning of the entire process of defining the position through composing the advertisement or job description.

The main nonobvious problem that comes with the two-stage approach is the possibility of a shift of standards once the microscope is brought out. This might mean that someone who didn't make the short-list would as a matter of fact have done well. It is quite a common situation. One contribution to coping with this problem is to avoid the common practice of sending out rejection notices immediately after the short-list has been identified. Instead, postpone the mailing until after the first meeting of the evaluators to rank the candidates on the short-list; and set the decision about sending out rejection letters as the last item on the agenda for that meeting. That reminds everyone of the possibility that some of the first-round rejects should now be reconsidered. Another way to reduce inconsistency is to use a process of explicit scoring and combining. Which brings us to the heart of the matter; how do we get from our data to a selection?

There are two general approaches; the global and the analytical. Most selection is done by forming a global (overall, holistic) impression from all the data. This approach is often quite adequate for identifying the best candidate (or the acceptability of the sole candidate). The problem is that it sometimes leads to bad results, and it is never very good at resolving close competitions. To avoid those problems, which are not usually foreseeable and are often undetected even when the process is complete (although they turn up on appeal), we should probably consistently use the analytical approach. In that approach we identify the main dimensions of interest, score each candidate on all of them, and then combine them in an agreed-on and defensible way.

The analytical approach is sometimes made so complex as to seem very objectionable to the evaluators (*bureaucratic* is a popular term). This is by no means necessary. If we start with the moment when the decision is made to advertise a position, our first task is to identify the language that is to be used in the job description, which will either be inserted in the advertisement or in the material mailed out to inquirers. That language should specify the basic set of *job-specific* criteria for merit, and the *institution-specific* criteria for worth (including the institutional mission statement, if any), although they may be phrased at a rather general level²³.

In fact, it is quite salutary to think of the advertisement (or job description) as incorporating the formula that will be used in subsequent assessment of applications, because it encourages being as specific as possible — without being too restrictive. The key tasks are: (a) identifying all characteristics (X1, X2, etc.) that are to count toward merit; and (b) deciding whether to say: "characteristic Xn will be *required*" or "characteristic X will be *preferred*." Remember that we will be held in violation of law, in many jurisdictions, if we consider seriously any candidate who does not meet the conditions we have said are required. In order to decide between those alternatives, we just keep asking two questions:

1. Will the result be so broad that some highly qualified people that we want to apply will refrain because they see themselves as having to compete with too broad a field to make it worth undertaking the psychic and logistic effort of applying?; and
2. Might the specification be too narrow and, hence, preclude an applicant that is really wanted, who didn't quite fit what has now been laid down?

For methodological reasons, it is highly desirable at this stage to keep the number of criteria to somewhere in the range of 4-7, if necessary by grouping them. Fewer criteria means they will not serve to resolve disagreement²⁴; more makes the whole process of scoring and combining too complex.

The advertisement (etc.) should contain some indication of the relative importance of the different criteria. While we should avoid any elaborate quantitative exercise here, as numerical weights are very hard to justify, and so setting them wastes a great deal of time, we can afford and do need to talk about allocating 5 or 10 points to various achievements. (Ratios of 2 to 1, and point allocations across a 10-point scale, especially if agreed to after discussion, can be made with reasonable objectivity.) The key point is to begin by identifying the necessary level of achievement on those criteria that you have identified as essential (e.g., must have a current credential, a major in mathematics, three years experience of teaching in this state, etc.), and then mention characteristics that would be desirable, possibly grouping them into more and less important²⁵.

Once we have constructed the job description, and the criteria/weights embodied in it, we simply score the candidates on them. Details of one reasonably thorough but not impossibly complex procedure for dealing with a substantial number of applicants are as follows. It's a modest refinement of the procedure suggested earlier.

1. At the first meeting, to avoid copying all applications for all raters, they are passed around and each rater classifies each candidate as no-way (fails to qualify on the required criteria) or possible or promising. The group now compares notes and discusses disagreements to identify: (i) the qualified list of candidates vs. the no-way group, who can be notified at this point; (ii) the short list of promising candidates (stars). In both cases, some further checks may be necessary to handle some queries in the second column of the

rating sheets (discussed earlier). They also discuss any problems that they see emerging from this first scan.

2. Copies of the qualified applications can be made for each panelist, which entails delay and expense, or the dossiers can be placed in a central but secure place (access via the principal's secretary or the librarian, for example) so that everyone can read these applications at leisure. If the applications are brief, all this can be done at a continuation of the first meeting, with someone working the phone to get the answer to any outstanding questions. The task now is to allocate the predetermined points for nonessential and overruled achievements (and worth considerations), in order to identify the short-list. All panelists should determine their own ratings before discussion begins, unless they spot something which raises a new issue of principle. Discussion now takes place to identify and rank the 2-5 candidates who will make up the short-list at this point. Particular attention should be paid to *worth* considerations, which can never be fully thought out in advance of seeing applications.

3. The group now has to decide as to the further investigations they feel are still required — and with respect to which candidates — including follow-up with referees, verification of credentials, phone interview, personal interview, proxy interview, or tests. And it has to be decided who will do this, and how they will report the results to the panel. It may or may not be necessary to have another meeting before inviting anyone in for an interview or running the phone interviews. It needs to be quite clear at the end of this discussion just what has been decided, and it is good practice for the chair to draw up a memorandum stating what has been agreed, partly because people's memories are not reliable, and partly to provide an audit trail in case of appeal. Doing so also facilitates ensuring consistency and steady improvement in the school's process of personnel selection.

Notes

1. Ranking is the process of ordering items by relative merit; an example is determining the medalists in a swimming meet.

2. Grading is the process of assigning items to a set of categories which are (i) defined in evaluative terms; (ii) independently defined — that is, they are not just defined in terms of each other, but in terms of some other concepts or measurements; (iii) ranked by merit; (iv) normally exhaustive of some dimension of merit. Although the set of categories is ranked, the items within any one category are not ranked. The usual grades A-F, as used in marking academic work, meet these conditions. Determining whether a solitary candidate for a teaching job is good enough to appoint is an example of grading on a pass/fail basis. In the swimming meet example, grading is involved in the decision as to whether or not someone meets the standards for membership of the Olympic team.

3. It is rare for a typical tax-funded school to develop one of these with any specific content (one often sees general motherhood-and-apple-pie statements).

4. Note that the *duties of a teacher* involve a great deal more than the duty of teaching. One version of the general duties list is in "Evaluating Teachers as Professionals," in *Teacher Evaluation: Six Prescriptions for Success*. J. and S. Popham, (Eds.) ASCD, 1988. A later version, substantially improved, can be obtained from the author at PO Box 69, Pt. Reyes, CA 94956 by sending a stamped, self-addressed envelope.

5. Recent and hence future performance may be radically different from earlier performance for a number of reasons: burnout, alcoholism, drug-use, changed interests, marriage or financial problems, or major problems with parent, children, or business. Reading this list should convince one that

these are not just problems with the older teacher; they are problems with people in general. One must not act as if they do not occur, and that means one must look very carefully at recent performance.

6. Here we are getting into the second topic (How should we go about the process of selection?), but we need to say something about it immediately.

7. It's convenient to say that worth comes only through these further contributions although in its literal sense worth comes from performance of duties, too. We use the term "value" to refer to the common currency, so $\text{Value}^2 = \text{Merit} + \text{Worth}$. The precise relationship is $\text{Value} = \text{Worth}$ from doing the job meritoriously + Worth from doing or being other things besides a meritorious worker within the official job slot.

8. Encouraging women to go into an area where they have talents, even when it is contrary to the common mores, is in general a contribution to the society, which needs the best people performing each job. (In this specific example, of course, the pay-off is especially crucial because of the shortage of highly qualified science and math teachers.)

9. It should also be noted that the cases of gender-weighting described are not being justified by appeal to compensatory justice. That consideration applies sometimes, as the courts acknowledge, but rarely to those who were not directly affected by the injustice. However, there is a sense in which what has been described is the legitimate descendant of compensatory justice. Periods of discrimination lead to a need for some degree of gender rebalancing, not for reasons of symmetry or compensation, but just because the result of the discrimination is empty roles of exemplar, adviser, and counselor.

10. There are other ramifications of the Case of the Swimming Pool Bribe; for example, once the principal has told the controversial teacher that his or her rights are completely safe, should the teacher resign voluntarily? Should the principal ask the teacher to join with the principal in a search for a better job for the teacher, provided that the search is done not under threat of dismissal even if such a job does not materialize?

11. Re-advertising is not a cure-all, because it will bring in some new candidates whose special qualities prompt a new allocation of Worth points. About the best one can do is to try very hard to identify the more obvious dimensions of worth, and also indicate that others may emerge in particular cases. This tells candidates that they should err in the direction of inclusion of virtues of possible value.

12. This is of course connected with the fact that very little attention is paid to content mastery in the evaluation of teachers who are already on the staff, a situation that occurs partly because principals feel it inappropriate to judge subject-matter mastery of those in fields other than their own. (I base this and other claims about current and recent practice on the comments and written examinations of approximately two thousand school administrators from all over the United States, who have taken my Nova University doctoral seminars on evaluation, and on similar comments from a much smaller number in Australia. The homogeneity of their reports is remarkable, and is not inconsistent with the majority of reports on these matters.)

13. In the duties list, these are separated into process management (discipline, question control, etc.), progress management (getting through the curriculum in the time available), and crisis management (what to do with fire, flood, and fits).

14. A detailed defense of this claim is provided in, for example, "Validity in Personnel Evaluation," *Journal of Personnel Evaluation in Education*, vol. 1, no. 1, 1987, Kluwer Academic Publishers, Boston, pp. 9-23.

15. We only need to be sure enough of your preliminary classification so that we can say: "Even if the best results came in from our further inquiries, this candidate could not compete with the several we have already verified."

16. If there are only one or two clear leaders, identify a second-string group of about 1-3 immediately, so that if the leaders decline or disappoint, you already know who to line up without a further meeting. Include them as part of what you refer to as the short list, as they are potential interviewees. The true cost of interviews is high (travel plus time away from normal work for panel and candidates), so it's very bad practice to set up interviews for a whole short list of four or five when

there are one or two clear leaders on paper. It's common to find that the research on applicants is done superficially, in the belief that it's better to interview a wide range of candidates. This is inefficient; better research often leads to discarding or promoting, and a much shorter short list. For reasons we'll expand on in a moment, the interview is a highly unreliable primary selection device; it's only good for verification and resolving residual questions.

17. Those who are concerned about the validity of student ratings might want to look at "The Validity of Student Ratings," *Instructional Evaluation*, Summer, 1988.

18. "Exhibiting enthusiasm" is not established by exhibiting a show of enthusiasm. One expects to hear details of recent work and study in the area, and of well thought-out perspectives on it.

19. Some search of the relevant managerial literature has failed to turn up good studies of this. The examples quoted are from personal experience in working for administrators in higher education who became curious about how we could have admitted or selected people that later turned out to require dismissal.

20. A combination of the two makes possible the use of evaluator with special knowledge who could not otherwise take part (e.g., a subject matter specialist when there is not one at the school site).

21. One can arrange a code of signals that will work even during a face-to-face interview. The following system is said to be practiced in the Army: one finger laid over the edge of the table means that one is satisfied that the interviewee has made it (either to the short list, or to the top of the short list, depending on which is at stake); two fingers means that one is satisfied that the interviewee should be ruled out from further consideration.

22. The absence of evidence that these competencies are present should be treated as equivalent to their absence, if they were clearly specified as essential in the advertisement or material sent to all applicants.

23. In addition, there are the commonly accepted "duties of a teacher," mentioned earlier. These are not repeated in every advertisement, but are part of the basis for the judgment of merit.

24. If we just advertise for "a science teacher for upper secondary," this suggests that we have two criteria: subject matter knowledge and teaching ability at that level. But *teaching ability* is too vague for good rating reliability. It's more useful to break it out into the four categories listed previously: testing skills, management skills, A-V/computer skills (to the extent relevant), and educational materials design. To these we should add the relevant personal qualities listed, and the worth factors that are most important to us, for a total of 7. If we go up to 10 criteria, we get an increased agreement about the performance on each, but the relative weightings then assume great importance and are quite hard to establish with any validity.

25. We might include some mention of how we will weight the "overruns" on the required dimensions, e.g., six years of teaching instead of three; while getting into any detail here is not required, the usual language is: "A minimum of three years teaching experience is required, but more would be desirable." Or "... more would be of some value," or "very desirable." We do not have to mention that we will not give extra points above, say, eight years experience — that's a level of detail beyond the requirement of fair advertising — but our committee should have settled that point. A good general procedure here, assuming we have 4-12 main criteria, is to allocate a certain number of points up to 10 for overruns (bonus performance) on each criterion (a maximum of 10 for the "very important" and 5 for the "desirable"), to be allocated after the candidate has passed on the required minima. For example, three years teaching experience is essential, so we know that all the candidates we are looking at after cutting out the unqualified have that much; treat that as the baseline or arbitrary zero for further evaluation. Now we might decide that overruns on teaching experience can score up to 5 points (one point a year); on the other hand, we might also allocate 5 points to "interest and skill in coaching basketball", which isn't one of the essential characteristics but was mentioned in the job description. Again, a teaching diploma is essential, but we allow up to 10 points for further degrees, just as we do for "proven record in management of computer resources," which is also nonessential. Thus, a criterion may be essential although overruns on it are of zero or lesser value than performance on a criterion that is not essential. The one tricky bit here is deciding how much to allocate to *worth*. Worth is the "everything else of value" category, and in part its weighting depends

on how thoroughly you built what would otherwise be worth considerations into the job description. The only arrangement that seems to make sense is to say that it cannot be more than the sum of all points allocated to merit (anywhere near that point and one should readvertise), and it will usually be hard to get it above the level of one relevant merit consideration (i.e., 5 or 10 points). However, it's no use putting a sharp limit on it; there are exceptional candidates for particular jobs where it may come close to matching the importance of everything else. In such cases, substantial discussion will be required to settle on a figure.

Note that the choice of 5 and 10 points is not simply arbitrary; using 50 and 100 points won't work, although 3/6 or 5/15 is fairly satisfactory. The figure is constrained by the number of dimensions, and if we are talking about 4-12 dimensions, 5 and 10 gives a reasonable match to our intuitions.

6. Assistance and Assessment for Beginning Teachers

Kenneth D. Peterson

First- and second-year teachers contribute a great deal to school systems, including enthusiasm, energy, and the innovations of recent training. However, along with these positive features come areas of concern. New teachers generally show gaps in their training, have unrealistic attitudes about work, and face demanding performance requirements. They must adjust to complex and often puzzling sets of rewards, roles, relationships, and expectations of being a teacher. Beginners lack self-assurance. A brief track record contributes to concerns of many audiences about the competence of new teachers. School systems rarely treat beginning teachers differently from veterans, except at times to furnish poorer working conditions. All of these concerns mean that beginning teachers need provisions beyond those given to experienced teachers. This chapter will discuss teaching as it looks to beginners, specific activities for successful teacher induction, and increased evaluation that informs and reassures teachers and their audiences.

Examination of the conditions and concerns of beginning teachers leads to two major goals for educators. The first goal is to develop effective teachers who take initiative, are self-regulating, and contribute toward the professional collective. The second goal is to retain promising performers. While the dynamics of beginning teacher dropout are not entirely documented, it is clear that a retention problem exists both for individuals and the profession (Heyns, 1988).

Induction and evaluation programs can be designed specifically to address the needs and goals of new teachers. The pay-offs for such programs go beyond development and retention of good teachers to include outcomes of better information about what teachers are doing, dissemination of good practice, improved socialization, and upgrading of evaluation for all teachers. Because beginning teacher needs reside in a population that is politically weak and personally transitory, there are few effective advocates for increased assistance and assessment for beginning teachers. Fortunately, there is a rapidly growing

AUTHOR'S NOTE: Acknowledgment and appreciation are expressed to Leslie Huling-Austin (Southwest Texas State University) and John M. Johnston (University of Wisconsin, Milwaukee) for their considerable information and perspective for this chapter.

number of programs that address beginning teacher needs and perspectives (Griffin & Milles, 1987).

Beginning Teacher Needs and Perspectives

A useful picture is emerging from the research literature on concerns and views of new teachers. Johnston and Ryan (1983) divided the literature on new teachers into reflective interpretations, advice about the first year of teaching for beginners and others, scholarly essays on what ought to be, and self-reports. Veenman (1984) summarized more than two decades of research reports concerning beginning teachers; he included a catalog of perceived problems, person-specific difficulties, and forms of planned support. Huling-Austin (1987) described teacher induction as a part of the bigger picture of the entire teaching profession. She reviewed difficulties experienced by newcomers, gave directions for needed practices and research, and linked induction to larger educational issues on attracting and retaining a high quality teaching force. These authors, and others, point to three particular areas of beginning teacher needs: professional growth, sociological development, and evaluation.

Professional Needs

Beginning teachers have distinctive and predictable needs for professional assistance. Most immediate is familiarity with the students and materials with which they work. Previous experiences in student teaching rarely match the actual pupils and conditions of the first-year assignment. Second, few preservice programs are comprehensive, but rather leave performance areas neglected. Information, techniques, and materials must be available when the beginner encounters demands and realities of the classroom and school. Next, new teachers rarely have learned how to gauge the effects of their teaching on students and how to see the connections between what they do as teachers and what their students learn. Most beginners need help with perspective on judging the effectiveness of their efforts. Fourth, first-year teachers may be extreme in their practice, employing limited instructional alternatives or unrealistic ideals calling for high labor intensity and extreme organization. Both of these situations require professional assistance. Finally, beginners require mundane models of how to do the perfunctory tasks of teaching. Paperwork, traffic management, and supply logistics exemplify learned procedures that make the difference between effectiveness and frustration in classrooms.

Johnston and Ryan (1983) identified pupil evaluation, planning, and organization as specific common competency gaps in beginning teachers. In addition, these authors described problems of pupil motivation and discipline that chronically plague newcomers. Chief problems listed by Veenman (1984) were discipline, motivation, individual differences, student work assessment, parents, or-

ganization, materials, and individual student problems. The predictability of these professional needs suggests that systematic support provisions should be anticipated.

In addition to specific professional performance needs, there are additional areas of support that may seem at first to be tangential to classroom performance. Beginners need to learn skills for working with adults: other teachers, administrators, parents, and schoolboard members. New teachers have not been recruited to work with adults, not many teacher-education programs train for interacting with adults, and few expectations may have been discussed at job interviews. Second, such topics as time management, changing relationships, and personal life concerns require attention for optimal professional performance. Personal life problems (such as living arrangements, personal finances) are remarkably consistent and important in shaping a new teacher's professional thoughts, values, and practices.

Socialization

Becoming a teacher is more than just acquiring information and teaching it to others. It also means functioning in a complex human interactive environment. Specifically, beginners must learn to operate with the roles, relationships, rewards, sanctions, and expectations of a variety of perspectives. A teacher is not just a collection of feelings, thoughts, and values held when he or she decided to enter the profession. Personal social norms and sentiments profoundly and predictably change to conform to a schoolteacher's perspective. The social perspective, in turn, will inform and limit professional performance.

New teachers rarely are treated specially and with a long-term perspective. The traditional induction is to allow teachers to sink or swim: learn to cope quickly or deserve the weeding out that results from early career dissatisfaction. Sociologists (e.g., Lortie, 1975) have described the veterans' message to perplexed neophytes as "welcome to the real world; we went through it, and so shall you." Given this lack of support, beginners must rely upon personal frameworks, achievements, intuition, and recollections of their own school experiences to cope with first-year teaching challenges (Blase & Greenfield, 1982). What develops in teachers is a sense of rugged individualism, when in fact the majority of problems are shared in common and a collective perspective among teachers can be supportive and useful. Lortie described this socialization as the Robinson Crusoe approach, in which the key sources of teacher growth are ad hoc solutions to personal errors.

Because other direct measures of efficacy are hard to come by, beginning teachers have strong needs to be accepted and liked. The complexities of classroom events interact with new-teacher feelings to produce pendulum swings between overtolerance and overreaction. In addition, beginning teachers must deal with difficult adjustments in personal-professional self-concept. Wright and Tuska (1968) found that the first year of teaching resulted in beginners seeing

themselves as less happy and inspiring, and more impulsive, controlling, and blaming than when they entered teaching. Ligana (1970) represented a "curve of disenchantment" during the first year during which idealistic and uncertain professionals face a period of self-reorientation.

Sociologists have found few effective rewards for teachers (Lortie, 1975). The strongest sources of acknowledgement for class control, pupil work compliance, and respect and affection satisfy only in the short term. Additional erratic rewards from the gratifying graduate, spectacular student case, and "good day" are not equal to the complex and profound demands of teaching. Instead, teachers need reward structures that go beyond the superficial: esteem based on concrete evidence of some configuration of pupil progress, peer recognition, systematic client (student and parent) approval, professional growth, and mastery of subject matter. However, veteran teachers are not comfortable with increased rewards. Ironically, in present-teacher culture it is not considered good form to go-public with exemplary practice. The very scarcity of rewards for teachers has produced a teacher society in which members discourage one another from the few available rewards.

Researchers in the area of teacher induction point to too few programs to assist beginners with the above concerns of adapting, coping, and working together. While progress is being made in some locations, most beginners actually have LESS support than veterans. Seniority assignment systems too often mean that beginners get castoff classes and most of the classroom sharing and floating assignments. Colleague acceptance and inclusion most often is withheld until the veterans get their agendas taken care of. Isolation, a problem for most teachers, is more pronounced with beginners who do not get to work with familiar people. One prominent deficiency for teachers is that they get very little feedback about their work.

Evaluation

Beginning teachers face three problems of evaluation that are shared by veterans and one problem that is unique. First, evaluation practices generally are poor, inaccurate, and provide limited feedback (Medley & Coker, 1987). Second, most school systems create a conflicting role for the principal as both supporter and judge. Third, formative and summative evaluation are done at the same time by the same person. The unique problem is that feedback given to beginning teachers differs from that of university supervisors (Cook & Richards, 1972).

While it is apparent that the evaluation of new teachers needs to be improved and increased, great care must be taken in making changes. Beginners should be given protection, even some privacy, in early evaluations in order to manage the level of threat. Early-career, high-stakes evaluation, as in some new career ladder programs, encourages conservatism, self-protection, and penalties for the most diligent and conscientious.

Just as beginning teacher instructional needs may be seen as distinct from veterans, so are attitudes about evaluation different. Beginning teachers want the authoritative reassurance provided from evaluation. They believe in assessment data and expect it. As expressed by one disappointed two-year veteran, "I thought I would be told how good a job I was doing. I don't know who I actually expected to tell me — some combination of my college supervisor, the principal, and my mother." Bad evaluation experiences and socialization from skeptical veterans create professionals who do not support teacher evaluation.

Beginning teachers can be developed and retained through induction and evaluation systems which attend to the characteristics and needs identified in this section. Provisions for successful teacher induction will be discussed next; effective evaluation of beginning teachers will be presented in a later section.

Assistance for Beginning Teachers

The major goals for a teacher induction program include additional skills and knowledge, professional role development, a personalized transition, dealing with isolation and stress, reduction of common problems, integration into school community, and retention (Huling-Austin, 1987; Johnston, 1985; Veenman, 1984). Additional objectives may focus on topics of teacher development such as career stages and formative evaluation for career-long development (Brooks, 1987).

Teacher induction assistance can be targeted to the specific needs and characteristics described earlier (Schlecty, 1985). Table 6.1 presents eight such provisions, which are discussed below.

Provisions for Assistance

Orientation sessions with introductions enable newcomers to become familiar with district locations, resources, and provisions. Staged information (i.e., before school year, midyear, end of year) is important to not overwhelm at first, but to continue orientation over several sessions.

Table 6.1 Provisions for Beginning Teacher Support

Orientation to tasks and district
Appropriate job assignment
Supportive working conditions
Inservice education
Mentors
Visitation opportunities
Support groups
Evaluation program

Appropriate teaching assignments are very important to beginners. Roth (1986) pointed out that 12.4% of new teachers in his large-scale study were assigned to areas for which they were not certified. Additional numbers are assigned to classes or courses for which the teacher preparation is weak. Reasonable assignments, even somewhat sheltered in terms of student numbers and difficulty, greatly assist the beginner. This practice may call for a change in seniority assignments. The reality of public schools is that the toughest assignments are too much for most newcomers. The attitude of "I went through it" held by many veteran teachers must be changed to "let's give you time to become as good as we are."

Supportive working conditions for beginning teachers can include a desirable and accessible room; many newcomers are not given a homebase room. Sufficient and available materials lessen the demands of teaching for less resourceful beginners. Extracurricular duties can be limited, and planning time protected.

Inservice education can at first be targeted to specific, local, in-place programs. The goal is familiarity and expertise with the immediate curriculum and instruction. Next comes more individualized diagnostic inservice to address gaps in preparation. Additional short-term specific courses provided by district or university, such as in planning or parent relations, address predictable needs. A discipline refresher course is needed by most new teachers once they have become familiar with the school population. Assistance with curriculum material centers needs to be available. These provisions can be followed by a long-term plan that blends into the inservice program for veterans.

Experienced teachers can be mentors for beginners (Driscoll, Peterson & Kauchak, 1985). Well-trained mentors serve as guides, consultants, and advocates; they should not be evaluators. Released time for mentors makes them available. Additional teachers who are not all-purpose mentors can be on call as consultants for specific topics, for example, discipline or reading instruction.

Visitation programs give beginners an opportunity to see good teaching practice and to get solutions to immediate concerns. The best visitation programs are not built on a vague identification of superior teachers, but on specific exemplary classroom provisions or strategies. Visitations require substitute time for the beginner, and brief discussions with the host teacher.

Beginning teachers can benefit from time spent with other newcomers. These sessions provide some orientation and inservice content, but their social value is considerable. Including new teachers from other districts is personally reassuring and professionally enlightening.

Assessment procedures can acknowledge early successes and reassure beginners that their preparation background has begun to pay off. Good assessment provides feedback on effectiveness, is responsive to the strengths of the teacher, and can be controlled in content and extent by the beginner.

Each of the above eight provisions for assistance of beginning teachers contributes support for good practice, desirable socialization, and accommodation for the significant personal challenges inherent in becoming a teacher. The

last provision, assessment (or evaluation), presents special opportunities for supporting new teachers which will be discussed in the next section.

Assessment for Beginning Teachers

Beginning teachers need and can use assessment to a greater degree than most veterans. This section will present teacher evaluation procedures which go considerably beyond typical practice. Such extensive efforts are justified for four reasons. First, the technical performance of teaching requires considerable information feedback, which is not available from casual student reaction, infrequent principal visits, haphazard parent comments, and teacher reputation. Second, sociological demands on teachers require authoritative reassurance provided by sound evaluation, increased rewards and acknowledgment, and lessening of the conflict placed on the principal as both supporter-leader and sole summative judge. Third, the political realities of teachers in this society mean that the public no longer takes the word of educators about quality, but expects teacher performance data and requests for financial support to be considered along with other demands, such as maintenance of highways, health care, and police protection. Fourth, developing good assessment practices in beginners is a way to reform evaluation for all teachers. Fortunately, initially positive attitudes about evaluation mean that novices are likely to support assessment activities more than are veterans.

Five recent innovations enhance the effectiveness of teacher evaluation for beginning or veteran practitioners: multiple data sources, variable sets of data, teacher dossiers, teacher control and involvement, and use of judges in addition to principals (Peterson, 1984). This section will discuss these assessment features and suggest uses for evaluation data.

Multiple Data Sources

Teachers can use performance data from a number of sources. No single line of evidence, even observation or principal reports, provides a total picture of what a teacher does (Peterson, 1987). Table 6.2 lists data sources which are useful to many teachers. Each of these lines of evidence provides some valuable information for beginning teachers, and those who assist them, to reassure and to shape and judge performance.

Useful assessment begins with information about student learning. Data about pupil gains on major goals and some specific objectives help to counter the endemic uncertainty of "Is anything happening?," which afflicts even veteran teachers (Lortie, 1975). Student achievement can be estimated with a combination of standardized and teacher-made tests given before and after instruction. Some approximate comparisons with other teachers gives useful perspective. While documenting absolute gains in a classroom is relatively easy, given many

Table 6.2 Data Sources for Assessment

Pupil achievement data
Pupil surveys
Parent surveys
Documentation of professional activity
Peer review of materials
Systematic observation
Teacher tests
Administrator reports
Other customized data

opportunities for testing and informal observation, producing student achievement data for summative or comparative teacher evaluation is difficult (Medley, Coker & Soar, 1984).

Pupil surveys are a valuable data source because teachers are especially concerned about student views of teaching and pupils are good reporters of many classroom conditions. Care must be taken in survey item selection so as not to create mere popularity contests, to avoid topics for which students are not good reporters (e.g., teacher subject matter knowledge), or to reward one teaching style. Recommended topics that indicate an opportunity to learn include presentation of new material, opportunity to practice and perform, reasonable control, and an overall global statement of teacher effectiveness. A preliminary survey midyear gives the teacher an opportunity to make midstream changes.

Parent surveys can reassure that an important audience recognizes that the teacher has created an opportunity to learn. In addition, surveys can indicate whether or not the teacher has provided parents with information needed to understand the class, to have confidence in student progress, and to support learning at home. Parents can report on sufficient challenge and whether or not the student seems to know what is expected in class. Surveys assist with communication and can provide satisfying compliments as well as suggestions.

Beginning teachers should be encouraged to keep records of their professional activity. Logs of classroom visits, inservice courses, training activities, and professional meetings all quickly add up to indicate preparation for effective practice. Without systematic documentation, opportunities for acknowledgment, reassurance, and building a cumulative case are lost.

Peer assessments are powerful influences on teacher practices and attitudes (McCarthy & Peterson, 1987). Not only is the professional knowledge that is communicated important, but the credibility and reassurance are significant to colleagues. Peer review of materials, done in the second or third year, can show the grasp of curriculum and program development expected of a beginning teacher. Peers are especially good at assessing appropriate challenge, range of materials, curriculum organization, and program completeness.

Systematic observation, with sophisticated procedures, observers, frameworks, and results, is rare and expensive (Medley, Coker & Soar, 1984). It can

address technical classroom performance such as opportunity to learn and equal opportunity to learn for high and low achievers, both sexes, and for ethnically different students. Observation also can check the cognitive levels of learning presented to students; a balance of higher and lower levels usually is appropriate. Finally, data and expectations on time allocation are needed by most beginners.

Three additional data sources may add to the overall picture of teacher performance. Teacher tests can provide assurance that the teacher knows his or her subject matter. The usefulness of standardized tests for new teachers is highly variable because practitioners differ so much in confidence of subject matter preparation and ability to take standardized tests. Administrator reports are reassuring, but usually are bunched at the upper range. Finally, customized data for unique talents or contributions are helpful for some teachers.

Extensive data gathering, as suggested in this section, may seem unnecessary and even prohibitive given short-term time and dollar costs. However, the long-term technical, sociological, and political gains are considerable. Information will be gathered from these sources in any event, but unsystematically and haphazardly. A teacher may overreact to the views of three or four parents, when a more representative sample gives another impression. While too much assessment is intrusive, a balance is needed to provide sufficient systematic data to permit realistic, accurate viewpoints and reactions. Individual teachers require different amounts of data. The first years of practice usually require more attention to assessment than subsequent years when teachers develop more complete views of performance.

Variable Data Sources and Teacher Dossiers

Just as teacher contributions and circumstances vary, so do the requirements for data sources. Each teacher may have a different configuration of evaluation lines of evidence. While many different sources are available, it must be kept in mind that no individual data source works well for all teachers. Data sources must be appropriate for the individual teacher.

Teacher data can be assembled in a dossier. A dossier is not a portfolio of work — better saved for a peer review — but a collection of data summaries limited to a dozen pages or so for ease of review. A collection of assessment data begun early and added to each year of practice becomes an important professional document.

Teacher Control and Involvement

Teacher control in evaluation is important (Peterson & Mitchell, 1985). Permitting teachers to call for data gathering places them in a politically important position. Second, allowing them to see results before deciding to place them in records is another protection. Teacher-controlled evaluation communicates a number of important messages. First, data gathering is expected and supported.

Second, there is variety in teacher performances. Third, teachers know their own situation best and can be trusted to contribute to assessment. Teachers' involvement in their own evaluation aids in the principal relationship because he or she is relieved from the role of sole data source and judge.

Judgments of Teacher Performance

Dossier evidence may be used in a number of ways. With the agreement of district, administrator, and teacher, additional evidence may be included in standard evaluations. For example, principal observations and reports can be enhanced with objective survey data, test results, reports of peers, and a complete record of professional activity. A second kind of judgment makes use of panels of peers or others assembled to react to the presented data (Peterson, 1988). Such reviews take sole-judge and validity pressure off the principal and reduce the isolation of individual teacher-performers by making more public the work on which judgment is based.

Effects of Evaluation

Evaluation for beginning teachers represents several kinds of interventions. First, there is a technical intervention of information about performance. This is formative when used by the teacher for improvement or by the district to shape inservice education programs. It is summative when used by the school for retention and tenure decisions. Both of these interventions are technical, or relate to performance of teacher tasks.

The effects of evaluation are sociological when they alter expectations, relationships, rewards, and roles. Evaluation data can serve for reassurance of the new teacher that expectations are met or exceeded. Evaluation affects roles and relationships; for example, when colleagues are involved in evaluation, rather than powerless subjects of it, significantly different concepts of teachers are created (Peterson, 1988).

The sociological and political outcomes of evaluation are arguably more than technical feedback. Beginners struggle with much more than feedback data, and few of them face negative summative judgments. However all new teachers adjust to expectations, face doubt and uncertainty, search for rewards and acknowledgment, and create a concept of being a teacher. Reassurance extends to colleagues, administrators, and becomes political to outside audiences of parents, lay public, and legislators who make decisions about allocation of scarce resources.

Costs for Beginning Teacher Assistance and Assessment

Costs for induction and evidence systems are substantial, but defensible. Personnel, materials, and time all are required for mentoring, data gathering, and visiting programs. As described by Johnston (1985), expenditures for effective induction offset replacement costs for teachers who otherwise would leave the system. Cynical attitudes and arrested professional development caused by haphazard, poor induction have a long term (although hidden) price. A faculty made reticent and conservative by early-career bad experiences is not as likely to respond as effectively to challenges as a group given initiative and confidence through effective inservice, useful data, and a sense of efficacy. A final consideration is that financing provisions for new teachers requires stability. Most induction and extended assessment systems face a lifetime of a few years, and close when seed money ends.

Discussion

The assistance and assessment efforts advocated in this chapter are targeted toward significant improvements in elementary and secondary schools. Provisions for support and data gathering are not made merely for a more pleasant or even effective first year, but to promote career-long attitudes and development. The emphasis in this chapter is not on the personal, idiosyncratic development of teachers, but rather is toward the thoughtful shaping of teachers who are able to control social forces which otherwise detract from good teaching.

In many ways the evaluation described in this chapter is good for any teacher. Proactive support, multiple data sources, feedback tied to inservice education, and teacher control all make sense at any career stage. The key is for beginners to develop sound data and attitudes. Without a successful foundation, teachers are remarkably poor consumers of evaluation, permitting disastrous practices and not demanding effective ones. Until significant improvements occur, teacher evaluation will remain something done TO workers, rather than something done BY professionals.

Lack of interested advocates in better support and evaluation of beginning teachers means a lower priority for such activities. The persons with a primary interest, the newcomers, lack power and perspective. It is the responsibility of veteran teachers, administrators, and policy makers to install effective assistance and assessment practices long known to make a difference for beginners. For example, the 20-year-old recommendation for reduced teaching loads for first-year teachers made by the National Association of Secondary School Principals (Hunt, 1968) is certainly not common practice. The need for farsighted action on behalf of beginning teachers remains. Veteran teachers in particular must be challenged to support short-term sacrifices of time, assignments, and resources for better and more professionally aggressive colleagues.

References

- Blase, J.J. & Greenfield, W. (1982). On the meaning of being a high school teacher: The beginning years. *The High School Journal*, 65, 263-271.
- Brooks, D.M. (1987). *Teacher induction: A new beginning*. Reston, VA: Association of Teacher Educators.
- Cook, M.A. & Richards, H.C. (1972). Dimensions of principal and supervisor ratings of teacher behavior. *Journal of Experimental Education*, 41, 11-14.
- Driscoll, A., Peterson, K. & Kauchak, D. (1985). Designing a mentor system for beginning teachers. *The Journal of Staff Development*, 6, 108-117.
- Griffin, G.G. & Milles, S. (Eds.) (1987). *The first years of teaching: Background papers and a proposal*. Chicago: University of Illinois.
- Heyns, B. (1988). Educational defectors. *Educational Researcher*, 17, 24-32.
- Huling-Austin, L. (1987). Teacher Induction. In D. Brooks (Ed.), *Teacher induction: A new beginning*. Reston, VA: Association of Teacher Educators.
- Hunt, D.W. (1968). Teacher induction: An opportunity and a responsibility. *The Bulletin of the National Association of Secondary School Principals*, 52, 130-135.
- Johnston, J.M. (1985). Teacher induction: Problems, roles and guidelines. In P. Burke & R. Heideman (Eds.), *Career-long teacher education* (pp. 194-222). Springfield, IL: Thomas.
- Johnson, J.M. & Ryan, K. (1983). Research on the beginning teacher: Implications for teacher education. In K. Howey & W. Gardner (Eds.), *The education of teachers: A look ahead* (pp. 136-162). New York: Longman.
- Ligana, J.W. (1970). *What happens to the attitudes of beginning teachers?* Danville, IL: Interstate Printers & Publishers.
- Lortie, D.C. (1975). *Schoolteacher: A Sociological Study*. Chicago: Univ. of Chicago Press.
- McCarthy, S. & Peterson, K.D. (1987). Peer review of materials for school teacher evaluation. *Journal of Personnel Evaluation in Education*, 1, 259-267.
- Medley, D.M. & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, 80, 242-247.
- Medley, D.M., Coker, H. & Soar, R. (1984). *Measurement-based evaluation of teacher performance*. New York: Longman.
- Peterson, K. (1984). Methodological problems in teacher evaluation. *Journal of Research and Development in Education*, 17, 62-70.
- Peterson, K.D. (1987). Multiple and variable lines of evidence in teacher evaluation. *American Educational Research Journal*, 24, 311-317.
- Peterson, K.D. (1988). Reliability of panel judgments for promotion in a school teacher career ladder system. *Journal of Research and Development in Education*, 21, 95-99.
- Peterson, K. & Mitchell, A. (1985). Teacher controlled evaluation in a career ladder program. *Educational Leadership*, 43, 44-47.
- Roth, R.A. (1986). Emergency certificates, misassignment of teachers, and other "dirty little secrets." *Phi Delta Kappan*, 67, 725-727.
- Schlecty, P. (1985). A framework for evaluating induction into teaching. *Journal of Teacher Education*, 36, 37-41.
- Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research*, 54, 143-178.
- Wright, B.D. & Tuska, S.A. (1968). From dream to life in the psychology of becoming a teacher. *School Review*, 76, 253-293.

7. Beyond Minimum Competence: Evaluation for Professional Development

Daniel L. Duke

Richard J. Stiggins

Evaluation can serve a variety of purposes in the lives of teachers. This chapter examines the extent to which teacher evaluation can promote professional growth beyond minimum competence. The chapter opens with a brief overview of professional development and why it is important. Next comes a review of the small body of research on teacher evaluation for professional development. This review identifies evaluation conditions associated with teacher growth. These conditions are grouped into three general categories: (1) characteristics of individuals; (2) characteristics of evaluation systems; and (3) characteristics of the environment in which evaluation takes place. The chapter concludes with discussions of policy issues and research needs related to evaluation for professional development.

Professional Development as a Focus for Teacher Evaluation

Teacher evaluation can serve two basic purposes — accountability and professional development (Stiggins & Duke, 1988). The accountability purpose — which is the focus of much of this volume — involves the collection of data to determine the extent to which teachers have achieved minimum acceptable levels of competence and prescribed areas or performance standards. Concern over accountability has tended to dominate the thoughts and actions of school officials charged with the responsibility for teacher evaluation.

AUTHOR'S NOTE: The authors wish to thank Marcella Dianda, Mary T. Rose, and Judith Little for their comments on early drafts of this chapter. The authors also wish to express their appreciation to Marvin Hempel and Frank Tassone for their help in shaping the ideas presented here. The preparation of this chapter was supported in part by the Office of Educational Research and Improvement, U.S. Department of Education under contract Number 400-86-0006. The contents do not necessarily reflect the views of the Department or any agency of the U.S. Government.

Interest in evaluation for the purpose of professional development, however, has been steadily increasing. Such evaluation has, as its central concern, the collection of data to assist teachers who are at least minimally competent in continuing to grow. Riegle (1987), in creating a conceptual framework for thinking about teacher development, identifies five general areas in which such growth can occur:

1. *Instructional development* emphasizes the development of skills involving instructional technology, microteaching, media, courses and curricula.
2. *Professional development* emphasizes growth of individual faculty in their professional roles.
3. *Organizational development* emphasizes the needs, priorities, and organization of the institution.
4. *Career development* emphasizes preparation for career advancement.
5. *Personal development* emphasizes life planning, interpersonal skills, and the growth of faculty as individuals.

Professional development, for present purposes, will encompass all five of Riegle's categories, not just category 2. *Professional development is defined as the process or processes by which minimally competent teachers achieve higher levels of professional competence and expand their understanding of self, role, context, and career.* This definition draws a clear distinction between the achievement of professional adequacy and the pursuit of professional excellence. The possibility is raised that conditions conducive to achieving professional adequacy may not be identical to those most likely to promote professional excellence. Before exploring the conditions that contribute to professional excellence, it is necessary to examine the potential value of professional development and the role of evaluation in facilitating its occurrence.

Is professional development really important? The pursuit of excellence, after all, may be a luxury school systems cannot afford. Is it not sufficient to hold teachers accountable for professional adequacy?

At least three justifications for professional development can be identified. First, many teachers appear to desire it. Teachers take advantage of a variety of opportunities for growth, including staff development workshops, peer observation, continuing education, and travel. That they may not see teacher evaluation as a source of professional development probably says more about how teacher-evaluation systems are designed and implemented than teachers' attitudes toward evaluation.

Both the National Education Association and the American Federation of Teachers have endorsed the use of teacher evaluation for formative purposes. Support by teacher organizations and individual teachers is prompting a number of school districts around the United States to consider adding a professional development component to their existing teacher evaluation systems.

Second, many states have begun to encourage or mandate continuing professional development for teachers¹. For example, Washington and Connecticut have requested that school districts propose new teacher evaluation systems with the intent that innovative systems for ongoing growth will emerge. The Texas Teacher Appraisal System assumes that "each teacher is capable of improving, regardless of the level of expertise or years of experience" (Texas Education Agency, 1986, p. 6). Oregon and North Carolina mandate that teachers annually must offer evidence of continuing professional development. California's Stull Bill requires that teachers biennially establish professional development goals. While these initiatives do not suggest that consensus exists on the exact nature of professional development, they do indicate that some states have recognized that teachers need new forms of assistance if they are to grow beyond minimum competence.

The third reason why professional development is important concerns the very nature of professions. One essential dimension of professions is a body of knowledge based upon systematic inquiry (Myers, 1973, pp. 23-24). Because new research causes these bodies of knowledge to change over time, it is expected that members of a profession — including that of teaching — will remain abreast of new developments. Failure to do so could place clients at considerable risk.

While few seem to dispute the value of professional development, some individuals may contend that mechanisms other than teacher evaluation are better suited to accomplish it. For example, the following mechanisms have been mentioned or utilized during the past decade as means for promoting professional development:

- State policy requiring continuing education units (CEU's)
- Job enrichment or promotion
- Career ladders
- Incentives and merit pay
- Released time for teachers
- Coaching of teachers
- Staff development
- Supervision
- Teacher transfer to a new school

We are not prepared to argue whether or not teacher evaluation is the *best* mechanism for promoting growth beyond minimum competence. Research comparing alternative approaches to professional development simply could not be found. What we can do in this chapter is examine the viability of teacher evaluation as *one* possible contributor to professional development. What teacher evaluation provides that may be missing in many of the other mechanisms is systematic feedback on performance. In the next section we look at the

contribution such feedback can make to the professional development of teachers.

Evaluation and Growth

How crucial is evaluation to growth? Our contention is that evaluation of some kind is extremely important, not just for the achievement of minimum competence, but for ongoing professional development. Human beings' potential for self-guided growth is limited by their cognitive structures, past experiences, and repertoires of skills (Knox, 1977, pp. 424-432). Once individuals have exhausted their own mental and emotional resources, they are unlikely to be motivated to grow without the intervention of some external impetus. Such an impetus can come in the form of a judgment or observation by a peer, supervisor, student, or parent. Evaluation feedback can provide the *challenge* found to be vital to stage growth (movement from one developmental level to another) by Thies-Sprinthall and Sprinthall (1987, p. 71).

While evaluation may stimulate teachers to grow beyond minimum competence, it also may have the opposite effect. Growth can be inhibited as a result of evaluation that is overly threatening, poorly conducted, or inadequately communicated. Upcoming sections examine what is known about conditions which are likely to spawn constructive, growth-producing teacher evaluation. These conditions can be divided into characteristics of (1) the individuals receiving and providing evaluation feedback, (2) the evaluation system, and (3) the environment or context within which the evaluation takes place. Before we address these factors, however, we need to comment on the knowledge base concerning growth-oriented teacher evaluation.

The Knowledge Base

In the handbook that preceded this volume, two chapters dealt with evaluation for the purpose of professional development. Iwanicki (1981) discussed the use of "contract plans," negotiated by teachers and their supervisors, as mechanisms for promoting teacher growth. Brock (1981) reviewed arguments for the use of evaluations of teaching performance (student ratings, classroom visits, audio/visual recordings) as a means of stimulating teacher development. While both chapters contributed greatly to the surprisingly small body of scholarship on the topic, neither relied primarily on empirical research into the use of evaluation as a basis for the professional development of elementary and secondary teachers.

When we conducted an ERIC search of written materials dealing with teacher evaluation for professional development, we obtained 95 citations for the period 1981-1988. Only 3 of these references actually dealt, however, with teacher growth beyond minimum performance standards, and none involved empirical investigation.

In terms of a research base on evaluation for the purpose of professional development, only a handful of studies are pertinent. Iwanicki (1982) developed and validated an instrument — the Teacher Evaluation Needs Identification Survey (TENIS) — for use in assessing the status of teacher evaluation programs. The 30 survey items for TENIS were derived from the Connecticut State Department of Education guidelines and criteria for effective teacher evaluation programs. To determine the construct validity of TENIS, a principal components factor analysis was conducted using the intercorrelation matrix for the 30 items based on the survey responses of 2,552 teachers. Eight factors were identified, including one dealing with "evaluation of personal improvement." This factor was defined as the "extent to which the teacher evaluation system takes a constructive approach in considering the personal needs of the teacher as well as the specific nature of the learning environment in which the teacher is involved" (p. 267). When Iwanicki utilized TENIS in a field study, he discovered that it successfully discriminated between two teacher evaluation systems — one a traditional system and the other a pilot program complying with new state guidelines for effective teacher evaluation.

A team of Rand researchers (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984) conducted a series of case studies of "effective" teacher evaluation practices that shed some light on the topic at hand. In their efforts to describe and understand the actual operation of four nationally recognized teacher evaluation systems, the researchers were able to identify practices associated with teacher accountability and practices intended to promote professional development. The researchers concluded, however, that both accountability and professional development purposes could not be served easily by the same teacher-evaluation system.

We reached a somewhat similar conclusion in a series of three studies conducted to determine the conditions under which teacher evaluation is most likely to foster the professional development of competent teachers (Stiggins & Duke, 1988). The first study, involving case studies of four teacher evaluation systems, revealed that "even highly formalized evaluation systems do not help teachers, either individually or collectively, to improve their skills" (p. 23). The second study required collecting narrative information from teachers who actually had experienced professional growth as a result of teacher evaluation. When the 33 narratives were analyzed, a variety of conditions associated with growth emerged. These conditions were grouped into five clusters: (1) teacher attributes; (2) evaluator attributes; (3) attributes of the information gathered on teaching performance; (4) attributes of the feedback on teaching performance; and (5) attributes of the evaluation context. Based on these clusters of conditions, the Teacher Evaluation Profile (TEP) Questionnaire was developed. The third study involved the pilot-testing and validation of the TEP with 450 teachers in five school districts. It was determined that the questionnaire provided high quality data regarding teacher perceptions of their evaluation experiences and that a strong relationship existed between specific attributes of an evaluation system

and the growth-oriented outcomes of that system as perceived by teachers (p. 115).

The conditions associated with teacher growth in our studies will be utilized to organize a discussion of evaluation for the purpose of professional development. The conditions — or attributes — can be divided into three general categories: (1) the characteristics of individuals, (2) the characteristics of evaluation systems, and (3) the characteristics of the context in which evaluation takes place. Individual characteristics are separated into characteristics of the teacher being evaluated and characteristics of the person undertaking the evaluation. Characteristics of the evaluation system include provisions for data collection and feedback. Contextual characteristics encompass general policies, purposes, and policies of school districts. Interestingly, these categories closely resemble Brock's (1981) three factors likely to influence the effectiveness of evaluation-based teacher development. These factors include contextual factors (organizational climate, resources, leadership, etc.), procedural factors (instruments used for data collection, use of other sources of feedback, etc.), and teacher factors (motivation, self-efficacy, etc.).

Characteristics of Individuals

Teachers. No matter how well-designed the evaluation system and how ample the support from supervisors and peers, teachers are unlikely to experience professional development if they are unable or unwilling to take advantage of opportunities for growth. In our case studies, we identified nine teacher characteristics that appeared to be linked to professional development (Stiggins & Duke, 1988, p. 94):

- Strong professional expectations,
- A positive orientation to risk taking,
- Openness to change,
- Willingness to experiment in class,
- Openness to criticism,
- Strong knowledge of technical aspects of teaching,
- Strong knowledge of subject matter, and
- Some positive prior experience with teacher evaluation.

Supporting research and scholarly judgment can be found in the literature on adult development and the professional development of teachers for many of these nine characteristics. Orientation to risk taking, openness to change, willingness to experiment, and openness to criticism are referred to particularly often.

Joyce and McKibbin (1982) found that teachers in the same school responded differently to the same opportunities for professional development. Upon further investigation they were able to identify five distinct teacher growth states:

1. *Omnivores* — These people actively use every available aspect of the formal and informal systems that are available to them. Their lives are rich with books, the performing arts, travel, sports, university courses, and the offerings of teacher centers and districts. They have found professional colleagues with whom they are close and interchange ideas. . . . They tend to be happy and self-actualizing people. Teaching has not jaded them nor has the rest of life.
2. *Active consumers* — People who take advantage of many (but not all) opportunities for growth and who occasionally initiate activities.
3. *Passive consumers* — People who are there when opportunity presents itself but who rarely seek or initiate new activities.
4. *The resistant* — People who are unlikely to seek out training unless it is in areas where they already feel successful.
5. *The withdrawn* — People who avoid virtually all growth-oriented activity.

Individuals who tend to take advantage of opportunities to growth often appear to welcome advice from others. They also capitalize on their own mistakes. Akin (1987, p. 45) notes that the "most common and powerful experience associated with the need to know, and thus with the start of important learning, is actual failure." Perhaps related to the ability to convert failures into constructive challenges is the tendency of growth-oriented individuals to avoid feelings of powerlessness. In a study of high achievers — or peak performers — Garfield (1986, p. 57) reported that no matter how rough things get, these individuals felt they could do something to improve the situation. Peak performers were characterized as constructively restless:

A task may appear impossibly difficult. Everyone else may be moping around. The peak performers can be found exploring possible solutions, something that might work. . . . They always have another idea to test, an extra possibility to explore (pp. 57-58).

Individuals who demonstrate a proclivity for continued growth seem to be more future-oriented than others. A controlled study of self-directed change by Winter, Griffith, and Kolb (1970, pp. 458-467) found that male college and university students who were more successful in achieving personal change goals displayed a greater ability to postulate future possibilities for themselves. Less successful students were deficient in this ability, suggesting that future possibilities were less salient for them than their present behavior. This latter group also appeared to be more confused or tentative about their present self, a syndrome referred to as *identity diffusion*.

Openness to change, risk taking, and feedback on performance do not come easily to many adults. Adults tend to be creatures of routine. Routines may be highly functional initially, but eventually they often become ruts. When Hannay and Chism (1988, p. 129) studied the differences between Ontario teachers who had and had not experienced professional growth as a result of a district teacher transfer program, they found that the latter group had "routinized their teaching and did not appear to differentiate between different learning situations, different grade levels, or different school settings." In addition, the researchers found that teachers who failed to perceive their transfer to another school as beneficial "evidenced a closed attitude toward change and sought the kind of stability at work [in their new school] that would permit them to continue their routine" (pp. 129-130).

Daniels and Zemelman (1985) write about the difficulties many teachers encounter when they participate in writing project workshops:

In an in-service workshop, we are not dealing with elementary school children who know they can write; we are dealing with grown-up adult teachers who, as often as not, believe they *can't*. As [Donald] Graves suggests, as people get older, they seem to accumulate more fears about writing and more reasons for avoiding it. . . . Given a roomful of colleagues for an audience and open choice of topics, many teachers (not all, but many) freeze up, resist, feel terribly uncomfortable, and might need days or weeks to thaw out, take some honest chances, and find an authentic voice in their work (p. 25).

Wlodkowski (1985, p. 45) has noted the important role of attitude in adult motivation. Attitude may be influenced by a variety of individual characteristics. A comprehensive review of research on goal setting and task performance indicated that the need for achievement and self-esteem may be the key variables in explaining individual differences in responsiveness to challenge (Locke, Shah, Saari, & Latham, 1981). Guskey (1988) conducted a study of 130 elementary and secondary teachers that found that teachers who were most receptive to the implementation of new instructional practices expressed a high level of personal efficacy, liked teaching, and felt confident about their teaching abilities. These are all attributes associated with greater instructional effectiveness. Ironically, the teachers who already may be quite effective are more likely to welcome new ideas and change than those who presumably have more to gain and less to lose as a result of professional development.

Facilitators of growth. A second set of items from our case studies encompassed characteristics of individuals — typically principals and peers — who were perceived to have facilitated teacher professional development. While the teacher characteristics listed in the preceding section were not found to be correlated with teacher perceptions of the value of their most recent evaluation experience, characteristics of evaluators — referred to henceforth as *growth facilitators* — tended to be highly correlated (Stiggins & Duke, 1988, pp.

105-106). The key characteristics of growth facilitators included the following (pp. 94-95).

- Credibility as a source of feedback
- Having a helper relationship to teacher
- Trustworthiness
- A nonthreatening interpersonal manner
- Patience
- Flexibility
- Strong knowledge of the technical aspects of teaching
- Capacity to model suggestions
- Familiarity with teacher's classroom/students
- Teaching experience
- Useful suggestion
- Persuasiveness of rationale for improvement

Of the 55 items in our Teacher Evaluation Profile Questionnaire, the 10 items that were most highly correlated with the teachers' ratings of quality and impact of teacher evaluation were directly associated with perceived characteristics of growth facilitators and the feedback delivered by that person (pp. 106-111):

1. Credibility of evaluator as a source of feedback ($r = .65$)
2. Quality of ideas contained in the feedback (.59)
3. Depth of information contained in the feedback (.58)
4. Persuasiveness of evaluator's rationale for improvement (.58)
5. Usefulness of evaluator's suggestion (.57)
6. Trustworthiness of evaluation (.56)
7. Helping relationship to teacher (.54)
8. Evaluator's capacity to model suggestions (.54)
9. Evaluator's technical knowledge of teaching (.53)
10. Evaluator's familiarity with teacher's classroom/students (.52)

These findings suggest that teacher growth — at least in the context of evaluation — is highly dependent on the perceived skill, integrity, and caring of those doing the evaluating. It is worth noting that the skills of the growth facilitator extend beyond technical expertise in evaluation to encompass general knowledge of teaching. Wlodkowski (1985, pp. 16-43) supports these findings when he identifies the cornerstone characteristics of individuals capable of motivating adults to learn. The characteristics include expertise, empathy, enthusiasm, and clarity. The aforementioned review of research on goal setting (Locke et al., 1981) reported that goal setting was most likely to lead to performance improvement when growth facilitators were supportive. Because it cannot be

assumed that the characteristics of effective growth facilitators are naturally abundant, extensive training is probably a prerequisite for those working with teachers to promote professional development.

Evidence is mounting that peers, as well as principals and supervisors, can play a key role in growth-oriented teacher evaluation (Glatthorn, 1987; Johnson & Johnson, 1987). It can be argued, of course, that fellow teachers are more likely than administrators to possess the empathy, understanding, and contextual awareness, not to mention subject matter knowledge, so crucial to the facilitation of teacher growth. That peer involvement has not been utilized more extensively is probably a function of three factors — the traditional isolation of teachers during the workday, the reluctance of colleagues to participate in the summative evaluation process, and the unwillingness of many administrators to share evaluation responsibilities. As teacher evaluation systems have expanded to include provisions for formative as well as summative evaluation and as opportunities for release time have increased, however, the prevalence of peer involvement in professional development has begun to grow. These observations imply that the success of growth-oriented teacher evaluation may be a function, in part, of systemic factors.

Characteristics of Systems

Evaluation process. Our case studies initially revealed that teachers perceive that certain aspects of their evaluation process play a role in professional development. This finding was confirmed during the validation of the Teacher Evaluation Profile Questionnaire (Stiggins & Duke, 1988, pp. 107-108). Two sets of characteristics of evaluation systems were identified as important — characteristics of evaluation procedures and characteristics of feedback to teachers.

Characteristics of evaluation procedures that were found to be correlated to the perceived quality and impact of the evaluation experience (based on teachers' perceptions that they had grown professionally) included the following items:

- Clarity of performance standards
- Extent to which teacher was made aware of performance standards
- Extent to which teacher endorses performance standards as appropriate for their classroom
- Use of classroom observations
- Examination of data on student achievement
- Number of formal observations
- Number of informal observations

While some of the elements of teacher evaluation procedures listed above are closely associated with the accountability purpose, it is important to note that

many teachers also feel these elements have the capacity to promote professional development. The key difference may be how these elements are applied in practice. Wise et al. (1984) found, for example, that performance criteria were useful for professional growth when they were applied in a flexible manner. Requiring all teachers to meet uniform specifications served the accountability purpose, but did little to promote growth.

Another interesting finding is that other sources of data besides formal and informal classroom observations may have a role to play in growth-oriented evaluation. In particular, teacher records, classroom documents such as lesson plans, and evidence of student achievement were perceived to be of value. This finding is supported by Peterson's (1987) research on the use of multiple data sources for teacher evaluation.

In addition, Haertel (1986) has delineated conditions for the valid use of student achievement data in the evaluation of teacher performance. When those conditions are satisfied, the assessment results promise to provide teachers with concrete evidence of the effects of their instruction on student learning. Such evidence can provide a strong basis for identifying professional development needs.

It is one thing to collect useful data on teaching performance and quite another matter to translate these data into meaningful feedback for teachers. Nine characteristics of feedback were identified in the case studies, and all nine were found to be correlated to the perceived quality and impact of teacher evaluation. These characteristics included:

1. Quality of ideas for improvement
2. Depth of information
3. Specificity of information
4. Account of information
5. Extent to which information was descriptive (rather than judgmental)
6. Timing of feedback to promote attention on the message
7. Extent to which feedback was linked to standards
8. Frequency of formal feedback
9. Frequency of informal feedback

The value of feedback for growth purposes was confirmed by Locke and his colleagues in their review on goal setting and task performance (1981). Where feedback was provided to show progress in relation to goals, individuals tended to experience greater success. While few would dispute the benefits of performance feedback, those familiar with the hectic pace of life in most schools acknowledge that finding time to provide quality feedback is difficult. Without support and encouragement from the central administration, principals and other growth facilitators are unlikely to devote the time necessary to ensure that

growth-oriented evaluation succeeds. This fact brings us to the final set of key variables.

Evaluation environments. Evaluation systems do not exist in vacuums. As pointed out in a series of case studies by Wise, Darling-Hammond, McLaughlin, and Bernstein (1984), an evaluation environment is very much a reflection of the school system and even the community in which it is located. In our studies, we found several characteristics of these environments that were correlated to teachers' perceived quality and impact of teacher evaluation (Stiggins & Duke, 1988, p. 108):

- Clarity of policy statements regarding teacher evaluation
- Extent to which a primary purpose of teachers evaluation is perceived to be teacher growth
- Time allocation for evaluation
- Extent to which labor relations have been tranquil
- Availability of time during the teaching day for professional development

The perceived quality of teacher evaluation varied across the five districts from which data were collected to validate the Teacher Evaluation Profile Questionnaire (Stiggins & Duke, 1988, pp. 111-114). Large differences across districts also were found in the strength of relationships between individual and evaluation system characteristics, on one hand, and perceived quality of teacher evaluation, on the other. Based on our findings, it appears that environmental characteristics can help predict the extent to which teacher evaluation is perceived to promote professional development for teachers.

To assist those working with and for school systems to improve growth-oriented teacher evaluation, we next turn to the implications of our review of research for policy making.

The Need for Policies that Promote Professional Development

Assuming that professional development and professional accountability are both desirable, a central policy question is whether both purposes can be served well within the same teacher evaluation system. McLaughlin and Pfeifer (1988, p. 69) have collected case study data on exemplary evaluation systems that presumably serve both purposes. Based on four years of research and development efforts in this area, we are less certain that one system simultaneously can ensure accountability and promote growth (Stiggins & Duke, 1988, pp. 131-134).

Part of the difference of opinion may be traced to different conceptions of growth. Teachers may grow toward greater competence, as represented by some set of performance standards. But few expand their repertoire of skills, develop

new perspectives on old problems, or become virtuoso performers in specific areas of teaching. This latter type of professional development may not be possible within the confines of evaluation systems serving the accountability purpose. It is fairly easy to understand why this should be the case. The reasons include lack of time and fear of risk taking. Each reason deserves further comment.

In many evaluation systems, policy has required teachers to be checked annually or biennially for accountability purposes. This process entails conferences between teachers and supervisors (often principals), classroom observations, and written reports. Teachers and administrators have told us that these activities consume so much time that there is little opportunity to attend to professional development. An important policy question which must be addressed in the following: Is the time required to conduct accountability evaluations of competent, experienced teachers every year or two an effective use of limited evaluation resources? Assuming that the vast majority of experienced teachers are reasonably competent (and we have no reason to believe otherwise), would it not be more productive to invest greater time and effort in helping teachers grow beyond competence?

A second reason why many teacher evaluation systems currently are not serving the purpose of professional development concerns the role of risk taking in growth. As suggested earlier, growth often entails a degree of risk. Evaluation systems that mix accountability for minimum standards and professional growth may present too much risk for the competent, secure teacher. First there is the risk of trying to tackle a challenging professional development goal. There is always the risk of failure, and the greater the challenge the greater the chance of failure. Then there is the risk of having an uninformed or untrained supervisor fail to make the critical distinction between a challenging goal and minimum standards, leading to a negative evaluation. If this were to occur, the teacher could face serious sanctions. These risks may lead competent teachers to play it safe, subjecting themselves only to evaluations based on standards they easily exceed and avoiding the potential pitfalls of growth-oriented evaluation. Unless these issues of time and risk are addressed in a systematic manner, we remain skeptical about the possibility that one evaluation system can serve both growth and accountability purposes.

The time is ripe for policy makers to rethink teacher evaluation, and there are numerous indications they are doing so. States like Washington and South Carolina and numerous districts in the U.S. and Canada are attempting to separate evaluation for accountability and evaluation for professional development. To do so, they have altered regulations so that competent tenured or continued contract teachers need not undergo annual summative evaluations of basic performance standards. Such evaluations have been shifted to every two, three, or in some cases four years. One example of a new policy comes from the Halifax (Nova Scotia) District School Board:

[The policy] clearly emphasizes the professional and cooperative nature of teacher evaluation. Supervision (formative evaluation), which normally occurs in three out of every four years, forms the basis of teacher self-improvement and development. Evaluation which is summative (i.e., results in a report which is used for employment purposes) normally occurs once every four years (Lowe, 1988, p. 2).

During the period between evaluations for accountability purposes, teachers and supervisors in these new systems undertake professional development activities. While the results of these activities *are* evaluated, the purpose of the evaluation is to provide formative feedback rather than a summative judgment. Eliminating annual evaluations for accountability frees school administrators to devote more time and energy to helping experienced teachers grow, thereby permitting them to function more as instructional leaders (Duke, 1987). In addition, by involving peers in the evaluation activities attending professional development, administrators can be freed to spend more time with teachers who are deficient in minimum performance standards.

We have worked with more than 30 school districts in nine states to develop separate evaluation systems for accountability and growth purposes. In the process, we have encountered a variety of issues and concerns that policy makers may need to address. These involve resources for growth, training, teacher participation, and options.

Professional development requires resources. Teachers need release time, tuition reimbursement, materials, and travel funds. Both teachers and growth facilitators may require special training. It is our feeling that some funds currently dedicated to staff development (where all teachers participate in the same training activities) could be shifted to professional development with no harmful effects. We believe that many staff development activities are relatively useless for large numbers of competent teachers and that the money for these activities could be put to better use supporting growth efforts by individuals. If few funds for professional development are available, it is probably unwise for policy makers to consider launching growth-oriented evaluation systems.

Training is another important concern. In our early efforts, we assumed that creating a new evaluation system and securing resources for professional development were sufficient. We quickly learned that most teachers and administrators do not normally think of challenging goals to pursue. Both groups require considerable training in goal-setting, reflection, and analysis. Part of this training should involve awareness-building activities designed to help teachers consider alternative ways to frame problems, recast assumptions as hypotheses, shift perspectives, and break routines.

Teachers clearly have a stake in professional development. It is only logical that they should participate, therefore, in planning evaluation systems to promote their own professional development. Many states where teachers collectively bargain mandate that decisions regarding teacher evaluation must be reached at the bargaining table. The projects with which we have worked that have fared the

best have been designed and monitored by teams of teachers, principals, and central office administrators.

A final set of issues concerns options. Should experienced teachers be free to opt out of professional development and simply participate annually in the accountability purposes? What kinds of growth activities fit under the rubric of professional development? Can a teacher choose to join a weight-loss program, for example, or pursue some other type of personal goal as part of his or her professional development? Must all professional development activities represent individual goals or may some be goals undertaken by groups of teachers?

These questions and issues are the kind that must be addressed by policy makers and professionals if evaluation systems for the purpose of professional development are to succeed. In addition, a determination must be made of how to handle teachers in the professional development phase who fail to meet expectations in basic performance standards. Should these individuals be released from their professional development commitments and placed on a plan of assistance? Should they be able to continue to work on their growth goals while undertaking remediation?

The Need for Further Research on Professional Development

Until recently there were relatively few teacher evaluation systems where it was possible to study evaluation for professional development apart from evaluation for accountability. As the number of growth-oriented teacher evaluation systems has grown, however, it is becoming possible to investigate a variety of important questions, including the following:

- Are competent, experienced teachers more likely to grow in systems that combine or separate summative and formative evaluation?
- What level of resources is needed to promote ongoing professional development?
- Are teachers who participate in professional development more or less effective instructionally?
- Are teachers who participate in professional development more or less likely to remain in the teaching profession?
- Do relations between teachers and administrators improve as a result of the introduction of evaluation for growth purposes?
- Can growth-oriented teacher evaluation be linked successfully to career ladders or other teacher incentive systems?

Besides questions such as these, there are two fundamental issues that require attention. First, we need to increase our general understanding of what it means to grow beyond minimum competence. Case studies documenting teachers' feelings and reflections as they tackle growth goals are crucial if we are to learn

more about the conditions that promote and inhibit professional development. Schon's (1987) work on reflection-in-practice provides one model of how such inquiry might be conducted.

A second issue involves the extent to which formal evaluation is critical to professional development. Comparisons of teachers engaged in growth-oriented evaluation and teachers participating in other kinds of growth-producing activities would be quite valuable. Do activities that entail some type of external evaluation constitute a more effective and/or efficient use of available resources? Is such evaluation more helpful when it comes at the beginning or at the culmination of a professional development project?

Conclusion

Research on growth-oriented teacher evaluation, though very limited at present, suggests that competent, experienced teachers may benefit from evaluation that is not driven by a set of basic performance standards and the desire for accountability, but which instead provides opportunities for individual awareness building, goal setting, and professional development. Certain characteristics of teacher and growth facilitators as well as evaluation systems and school systems have been found to be associated with high quality teacher evaluations for the purpose of professional growth. The more we can examine these characteristics and extend the existing research base, the better able we shall be to assist teachers in growing beyond competence.

Note

1. The Northwest Regional Educational Laboratory sponsors a program that provides technical assistance to states and districts desiring to add a professional development component to their teacher evaluation system. Projects have been undertaken in nine states.

References

- Akin, G. (1987). Varieties of managerial learning. *Organizational Dynamics*, 16, 36-48.
- Brock, S.C. (1981). Evaluation-based teacher development. In J. Millman (Ed.), *Handbook of teacher evaluation*. Beverly Hills, CA: Sage Publications.
- Daniels, H. & Zemelman, S. (1985). *A writing project*. Portsmouth, NH: Heinemann Educational Books.
- Duke, D.L. (1987). *School leadership and instructional improvement*. New York: Random House.
- Garfield, C. (1986). *Peak performers*. New York: Avon Books.
- Glatthorn, A.A. (1987). Cooperative professional development: Peer-centered options for teacher growth. *Educational Leadership*, 45, 31-35.
- Guskey, T.R. (1988). Teacher efficacy, self-concept, and attitudes toward the implementation of instructional innovation. *Teaching & Teacher Education*, 4, 63-69.
- Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational Evaluation and Policy Analysis*, 8(1), 45-60.

- Hannay, L., & Chism, N. (1988). The potential of teacher transfer in fostering professional development. *Journal of Curriculum and Supervision*, 3, 122-135.
- Iwanicki, E.F. (1981). Contract plans: A professional growth-oriented approach to evaluating teacher performance. In J. Millman (Ed.), *Handbook of teacher evaluation*. Beverly Hills, CA: Sage Publications.
- (1982). Development and validation of the teacher evaluation needs identification survey. *Educational and Psychological Measurement*, 42, 265-274.
- Johnson, D.W., & Johnson, R.T. (1987). Research shows the benefits of adult cooperation. *Educational Leadership*, 45, 27-30.
- Joyce, B., & McKibbin, M. (1982). Teacher growth states and school environments. *Educational Leadership*, 40, 36-41.
- Knox, A.B. (1977). *Adult development and learning*. San Francisco: Jossey-Bass.
- Locke, E.A., Shaw, K.N., Saari, L.M., & Latham, G.P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 90, 125-152.
- Lowe, A. (1988). *In-service materials on teacher supervision and evaluation*. Halifax, Nova Scotia, District School Board.
- McLaughlin, M.W., & Pfeifer, R.S. (1988). *Teacher evaluation: Improvement, accountability and effective learning*. New York: Teacher's College Press.
- Myers, D.A. (1973). *Teacher power*. Lexington, MA: Lexington Books.
- Peterson, K.D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal*, 24, 311-317.
- Riegler, R.P. (1987). Conceptions of faculty development. *Educational Theory*, 37, 53-59.
- Schon, D.A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.
- Stiggins, R.J., & Duke, D.L. (1988). *The case for commitment to teacher growth: Research on teacher evaluation*. Albany, NY: State University of New York Press.
- Texas Education Agency. (1986). *Teacher appraisal system: Teacher orientation manual*. Austin, TX: Texas Education Agency.
- Thies-Sprinthall, L., & Sprinthall, N.A. (1987). Experienced teachers: Agents for revitalization and renewal as mentors and teacher educators. *Journal of Education*, 169, 65-79.
- Winter, S.K., Griffith, J.C., & Kolb, D.A. (1970). Capacity for self-direction. In M.W. Miles & W.W. Charters, Jr. (Eds.), *Learning in social settings*. Boston: Allyn & Bacon, Inc.
- Wise, A.E., Darling-Hammond, L., & McLaughlin, M.W., & Bernstein, H.T. (1984). *Case studies for teacher evaluation: A study of effective practices*. Santa Monica, CA: The RAND Corporation.
- Wlodkowski, R.J. (1985). *Enhancing adult motivation to learn*. San Francisco: Jossey-Bass.

8. Evaluating Teachers for Career Awards and Merit Pay

Samuel B. Bacharach

Sharon C. Conley

Joseph B. Shedd

Introduction

A few years ago at a conference on merit pay and career ladders, one southeastern state commissioner of education extolled the virtues of the merit compensation system he was about to implement in his state. When questioned by one of the authors of this chapter about how teachers would be evaluated under this new system, he dismissed the question by maintaining that the most important thing was "to take action, and worry about the evaluation system when we get to it." In his self-assurance, the commissioner even chastised the academics on the panel, maintaining that we think too much and worry too much. Without much surprise, we must now report that his entire compensation system collapsed because of the failure to seriously consider issues of evaluation. It is in this context that we argue that before we can consider how to compensate teachers, we must consider how to evaluate them. A compensation system is only as good as the evaluation system that undergirds it.

The current period of reform has been focused strongly on issues of compensation. Merit pay and career ladders are two of the approaches tried, but they both have several major shortcomings. One of these is the need for a better career-evaluation system to underly the compensation plan. This chapter attempts to tease out some of the characteristics of such an improved teacher-evaluation program.

First, we discuss characteristics and concerns of merit-pay and career-ladder schemes. Next, we ask how we should evaluate these schemes in terms of the

AUTHORS' NOTE: The authors wish to thank Bruce Cooper and Naftaly Glasman for their helpful review of an early draft of this chapter.

standards underlying them. Should the standards be merely minimum, intensely competitive, or constantly developmental? Then, we briefly look at the nature of teachers' work, in order to determine what we should be evaluating. We arrive at a conceptualization of teachers as professionals with growing, developing decision-making skills. We categorize these skills according to the different types of roles and functions that teachers perform every day.

This model lends itself to a three-stage pattern of career development, and correspondingly, three levels of evaluation. First, teachers learn to apply basic skills. Second, they learn to adapt to the unexpected and integrate the various skill areas. Finally, they become innovators, sharing their ideas and becoming leaders.

What implications does this model have for the design of effective career development systems? This chapter concludes with some lessons learned from merit-pay and career-ladders systems, and suggests some principles when implementing evaluation systems for career development.

Merit Pay

Merit pay allocates special payments or pay increments to some teachers based in part on some form of evaluation of their performance. Such pay systems purportedly give school districts special incentives to strengthen evaluation and training procedures, and give teachers special incentives to pay attention to evaluative feedback and to seek help in improving their performance (Bacharach, Lipsky, & Shedd 1984; Calhoun & Protheroe, 1983; Hatry & Greiner, 1984). Unfortunately, dissatisfaction with evaluation procedures, friction between administrators and teachers, and friction among teachers themselves are the three main reasons why districts typically abandon such plans soon after they implement them.

Merit-evaluation procedures are most often criticized for their lack of objectivity, but as our comments on evaluation systems indicate, that criticism is misleading. The pursuit of ever more objective evaluation criteria — and standards for drawing objective distinctions between different levels of merit on any given criterion — is an absolute necessity when the allocation of pay raises hangs in the balance. But pursuit of those objectives merely reinforces the weaknesses in already weak teacher-evaluation procedures, by diverting the attention of evaluators from those aspects of teaching that are most difficult to master and least subject to objective assessment.

Even if accusations of subjectivity and bias could be overcome, merit-pay systems can cripple the open communication needs for constructive discussion of aspects of performance that might need improvement. Teachers can hardly be expected to volunteer potentially adverse information about their own performance to evaluators who have control over pay raises.

Merit pay also discourages cooperation and sharing of job knowledge among teachers, by forcing them to compete with each other for a fixed pot of money. Given the nature of school district financing, there must be some quota or limit on the number or size of merit increments available. Unlike employees in sales and certain entrepreneurial occupations — where merit-pay systems often do work effectively — individual teachers have no opportunity to finance their own merit increments by generating new money for their employer.

By weakening evaluation systems, discouraging communication between teachers and administrators, and further isolating teachers from each other, merit pay (at least as it is currently implemented) exacerbates what are already three of the most serious weaknesses of our public school systems.

Career Ladders

Another popular set of reforms are so-called career-ladder or master-teacher plans. Such plans typically create hierarchies among teachers and provide for promotion from one level to the next (Hart, 1987; Malen & Hart, 1987). Most plans provide for some reduction in the amount of classroom teaching as a teacher advances up the ladder, as he or she assumes responsibility for new functions such as course development, program evaluation, teacher training, and service as a mentor for beginning teachers.

Teachers often find the career-ladder concept attractive. The flat job structures of most school districts provide few opportunities for teachers to have a sense of career advancement or professional growth, and what opportunities these districts do provide involve leaving the classroom and entering administration. Career-ladder plans allow teachers to assume more responsibility for school-wide programs without having to leave teaching altogether.

However, there is nothing inherent in the existing compensation or job structures of public education that prevents districts from involving teachers in decision making or providing them with more developmental support. Specifying that a teacher must receive a *promotion* before he or she is eligible to perform new and different kinds of responsibilities merely limits the number of teachers who can perform them, creating a structural obstacle where none presently exists.

Specifying that particular duties belong to teachers at particular levels of the system virtually guarantees that advancement will be limited by quotas on the number of positions at each level, by turnover among people already in those positions, and by competition with others seeking such positions. Indeed, several states specify explicit quotas in their career-ladder legislation, thus undermining their claim that career advance will depend solely or even primarily on a person's individual qualifications.

Although intended to raise the status of teaching, such plans merely reinforce the myth that classroom teaching itself is less than a fully professional activity

and the corollary myth that the only real profession in education is administration. Teachers who have been promoted to higher steps in the job hierarchy may continue to perform some classroom teaching, but it is still their other (nonclassroom) duties that justify their higher status and pay.

All of these criticisms are a result of one feature of career-ladder proposals that is largely taken for granted — that is, the assumption that a system of career advancement necessarily requires that there be some formal connection between different sets of duties and different career levels. It is not surprising that many people would equate career advancement with climbing what is, in effect, a job ladder. Promotion from one job to the next is a typical, if not *the* typical, form of career advancement in most occupational settings (Bacharach, Conley & Shedd, 1986).

Yet we know that there are occupations where hierarchical career structures operate without strict delineation of what duties belong to each level (e.g., assistant, associate, and full professional in higher education, and apprentice, journeyman, and master status in many craft occupations). The assumption that there has to be a formal link between career levels and job responsibilities represents the primary reason that teachers and their unions have opposed the enactment of most career-ladder legislation.

Unfortunately, while obsessed with the way teachers are paid, many of the reformists did not focus enough on the role of the evaluation system behind these new compensation systems. This is evident from the dilemma of objectivity that we discussed above as one of the problems of merit-pay systems. For a career-ladder or merit-pay system to be successful, we need to specifically identify the goals of our evaluation system. Are we trying to purge low performers from the system, or are we trying to reward excellent performers, or are we trying to improve performance of *all* system members? To understand the various evaluation systems underlying merit pay and career ladders, we need to examine the way that the standards embedded in them are applied.

Education Reform and the Issue of Standards

Most reform strategies in education have been justified on the grounds that they would "raise standards" for teachers and thereby improve the quality of public education. However, reformers have actually advocated three different types of standards: minimum standards, competitive standards and developmental standards. Each type of standard is associated with a different strategy to reform education and each may be used as the basis for a system of teacher evaluation.

Minimum standards are closely allied to remedial approaches to evaluation (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). Minimum standards specify floors (levels of performance that teachers are expected to meet

and maintain). They are designed to keep from entering or weed out so-called incompetent teachers. While any profession needs to have some set of minimum standards (particularly in screening those who desire entry into that profession), basing a profession evaluation system exclusively on minimum standards can have serious pitfalls. Although minimum standards may be a useful first step, critics have argued that such a system fails to reward teaching excellence. An evaluation system must do more than assure basic skills and knowledge; it must encourage professionals to develop their skills to the highest potential.

Competitive standards are associated with reform strategies that regulate the allocation of rewards among teachers; the purpose is to provide incentives and rewards for some form of superior performance or for particular kinds of performance. Indeed, this approach has received primary emphasis in the current reform movement. Merit pay and most career ladders employ competitive standards, as they award incentives (e.g., a monetary award or a promotion) to teachers on a competitive basis. Even though the requirements for obtaining such incentives are sometimes stated in terms of a teacher's absolute level or quality of performance, the number of available rewards is almost invariably limited. When the absolute level is attained by more teachers than there are rewards, their relative scores (i.e., competitive standards) determine who will get which or any reward.

Some advocates of career ladders would take exception to our use of the competitive-standards label to classify these systems. After all, they have criticized merit-pay plans precisely because they do generate competition among teachers and undermine collegiality and cooperation among teachers (see, for example, Rosenholtz & Smylie, 1984). The purpose of career ladders, advocates maintain, is to provide *all* teachers with opportunities for growth in their careers. Unfortunately, whether by intention or accident, the typical career ladder fails to accomplish that. Some plans (like those implemented in California and Florida) set explicit quotas on the number of higher-level positions open to teachers. This creates a system with the same problem as when competitive standards are used to allocate scarce merit-pay rewards. (See above discussion.) In other words, by definition, only the best (those with the highest scores) can advance.

Other proposals make no mention of quotas, but incorporate a feature that virtually guarantees that implicit quotas will exist: they specify that different duties would be assigned to teachers at different levels of such a ladder. Advancement thus depends on the number of positions available at the next higher level, regardless of how many teachers might "objectively" meet the evaluation standards for promotion (Bacharach, Conley, & Shedd, 1986).

For these reasons, competitive standards are clearly inappropriate in an evaluation system designed to facilitate professional growth. Competitive standards simply encourage teachers to compete for a limited number of available promotions and monetary rewards (the implicit assumption is that only a few have the right stuff to be leaders), rather than providing all teachers with the means and incentive to achieve greater professional growth. If only the exceptional teacher

can earn the reward, then why should the average teacher even try? Because there are (by definition) more average teachers than exceptional ones, a system based on competitive standards is thus apt to result in more demotivation than motivation. Indeed, evidence now emerging from the private sector (Peters & Waterman, 1982) indicates that organizations that achieve excellence are those that expect excellence from all of their members, and have cultures and processes that constantly reinforce that expectation. The notion that leaders of such organizations would ever tolerate, much less endorse, a quota on excellence is inconsistent with everything we know about such organizations.

Developmental standards are designed to encourage the development of teachers's knowledge and skills by specifying skill areas or targets for improvement. Rather than relying on remedial (i.e., minimum) or competitive approaches, developmental standards emphasize teachers' capacity for improvement and growth. Such a model is based on the principle that every teacher needs improvement — not because he or she is incapable of fulfilling his or her responsibilities — but because there is always room for anyone to increase his or her knowledge and skills, and because growth itself is a basic need of those in professional occupations (Hall, 1976; Hackman & Oldham, 1980; Wise et al., 1984). A system based on developmental standards would require us to specify critical skills in teaching as well as the means by which those skills develop over time.

Indeed, during the past decade, there has been a growing consensus that a teacher-evaluation system in education should facilitate growth and development of teachers' skills over the course of their careers (Iwanicki, 1981; Burden, 1982; Darling-Hammond, 1986; Schlechty, 1989; Hunter, 1988; Christensen, Burke, Fessler, & Hagstrom, 1983). However, in our view, many of the current merit-pay and career-ladder proposals will fail precisely because they do not evaluate teachers for skill development and focus instead on specifying minimum performance levels for teachers (remedial standards) or else fostering teacher competition for scarce rewards (competitive standards).

In this context, we also need to have a clear understanding and consensus about *what we want evaluate*. Will it be the time spent (hours), effort (sweat), ability (aptitude), or output (student test scores)? Clearly if the goal of an evaluation system is to be developmental, it should focus on something that teachers have a real chance to develop — their professional skills.

Integrating Skill-Based, Developmental Standards into an Evaluation System Aimed at Career Development

It is possible to generate developmental standards for assessing the level of a teacher's skills. If that can be done, then we are closer to being able to transform a merit-pay or career-ladder system into a successful career development plan.

Roles	Functions		
	Planning	Implementing	Evaluating
Instructional	Instructional Planning	Instructional Implementation	Instructional Evaluation
Managerial	Managerial Planning	Managerial Implementation	Managerial Evaluation
Counselor	Counseling Planning	Counseling Implementation	Counseling Evaluation

Figure 8.1 Teacher Decision-Making Skill Matrix.

Such a plan would include specific criteria for achieving competence in a particular skill area. It would also include standards not as the minimum performance required for employment, but rather a set of targets for growth in the various criteria. Finally, it would include promotion requirements that represent some combination of standards met across various criteria. These requirements would be progressively more demanding for each successive promotion.

Many people have suggested taxonomies of important teaching skills toward which a career-development-evaluation system could be directed. Given the uncertainty that permeates the teaching environment (Metz, 1972; McPherson, 1972; Lortie, 1975; Lampert, 1985; Doyle, 1985; Jackson, 1968; Bacharach & Conley, 1989), we have suggested a framework in which teaching skills can be referenced to three teaching roles and three teaching functions (Bacharach, Conley, & Shedd, 1986).

According to the framework, the basic roles (or purposes) that teachers serve are those of instructor, counselor and manager; the basic functions (or processes) they perform are the planning, implementation (or execution), and evaluation of decisions and their consequences. Together, these two rudimentary decision categories allow us to identify nine basic role-function combinations, which can be said to be part of a teacher's decision-making skills. Figure 8.1 presents a decision skills matrix of these role-function combinations.

Based on teacher development patterns proposed by Burden (1982), three developmental career stages have been identified (Bacharach, Conley, & Shedd, 1987)¹. The stages that are specifically predicated on the assumption that teachers progress from acquiring a basic set of teaching practices to developing decision-making techniques that help them deal with the uncertainty and unpredictability inherent in their environment. Each stage is characterized by a specific type of skills acquisition and each represents a developmental standard for assessing teachers' decision-making skills in a system of teacher evaluation. Table 8.1 contains a brief description of the career stages.

Application. During the initial stage of their careers, teachers focus on acquiring basic skills and competencies. In terms of the skills matrix (Figure 8.1), they master the ability to apply skills in each of the nine cells in the matrix to their work in the classroom. For example, when a teacher develops a lesson plan, he

Table 8.1 Career Development Stages for Teacher Skill Acquisition

Stage	Feature
<i>Application</i>	The ability to use one's skills and knowledge to make decisions in basic teaching roles and functions (see matrix above). Basic Skills
<i>Adaptation</i>	The ability to adjust one's teaching decisions to reflect the diversity and unpredictability of the situation a teacher faces, may face, or already has faced in the classroom. Variety & Integration
<i>Leadership</i>	The ability to share one's adaptations with others. Innovation

or she demonstrates application of the basic skill of instructional planning. Teachers who are able to apply skills in each area of the matrix to their work in the classroom possess the minimum skills and knowledge to make decisions in each cell of the decision-making matrix. Further development in the profession is not so much a matter of mastering new bodies of knowledge or new skills, but of building upon and expanding those basic skills and competencies. This further growth and expansion is indicated by the teacher's ability to adapt these basic skills to classroom diversity and unpredictability (i.e., the adaptation stage).

Adaptation. Having previously acquired basic skills and competencies, in the second stage of development teachers learn to adjust (adapt) their basic skills to the unique teaching situations they confront. There are two features of this ability to adapt: variety and integration. *Variety* concerns the diversity of techniques a teacher has available to deal with various teaching situations. To adapt decisions to meet the different needs of individual students and to adjust to unexpected learning situations, teachers develop a variety of techniques for planning, implementing, and evaluating instructional, counseling, and managerial decisions (i.e., the cells of the matrix). *Integration* concerns the teachers' ability to combine his or her planning, implementation, and evaluation decisions, thereby reflecting awareness of the effects of specific decisions he or she has already made and anticipation of decisions he or she will have to make in the future. For example, teachers integrate managerial decisions when they observe class and individual behavior in order to identify needed changes in plans, and then adjust plans while in the midst of implementing them.

Leadership. Once teachers have mastered adaptation skills, the third stage is leadership, in which teachers develop the ability to share their adaptations and innovations with others. While leadership implies some form of on-going contact and effort with staff members outside one's own classroom, it does not imply or require any formal assignment outside the classroom. Becoming a mentor for other teachers is one form of leadership; developing a new classroom technique and sharing it with others is another important form of leadership. Teachers develop leadership skills in or across any of the cells in the matrix.

The skills matrix and the career development stages (i.e., application, adaptation, and leadership) provide a conceptual framework for teacher evaluation. As

such, they can be used to create a set of developmental standards that can be used to assess a teacher's level of professional skill. In the next section, we examine the implications of the model for evaluating teachers for career development.

Evaluation Principles for Teacher Career Development

Working from the model proposed above, four evaluation principles have been identified (Bacharach, et al. 1987). While the methods of specifying particular evaluation criteria and standards are beyond the scope of this chapter and are discussed elsewhere (Bacharach, Shedd, Malanowski, & Conley, 1984), these basic principles may serve as general guidelines for developing a system of evaluation for teacher career development that avoids the pitfalls of current merit pay and career ladder schemes².

Principle One: Skill-Based Versus Performance-Based Evaluation

The first principle involves focusing the evaluation system on assessing skills that are likely to contribute to effective performance rather than on measuring performance itself. While it is not possible to draw a precise line between evaluating performance and assessing skills (because skills must be largely inferred from performance), the developmental standards (i.e., application, adaptation, and leadership) presented above do not require evaluators to make direct judgments about the success or failure of a teacher's specific actions (i.e., his or her performance). Rather, they allow evaluators and teachers to discuss the possible utility of different teaching approaches, without implying that the specific approach a teacher may have taken was necessarily wrong. Additionally, they provide teachers with opportunities to discuss the reasoning behind their use of a particular approach: after all, it is virtually impossible to make judgments about a teacher's ability to integrate various decisions without such a discussion.

An advantage to focusing on skill-based evaluation rather than on performance-based evaluation is that skill-based evaluation gives at least some minimal assurance that even the poorest students will be afforded the opportunity of learning with a skilled teacher.

All too often, performance-based evaluations simply rate teachers on the basis of outcomes, rather than on the skills they have acquired. This is especially true for those merit-pay and career-ladder schemes aimed primarily at reward distribution instead of career development. This puts the teacher at an immense disadvantage. If teachers are evaluated on output-oriented measures of their performance (for example, student standardized test scores), the implicit assumption is that teacher skill is merely a *sufficient* condition for good performance. On the other hand, if teachers are evaluated on their skills directly, the

assumption is that teacher skill is a *necessary* condition for good performance. That is, the teacher is no longer behind the eight ball (i.e., solely accountable for student learning), because the evaluation system recognizes other factors antecedent to good or bad performance. Obviously there must be a balance between output- and skill-oriented systems of evaluation. Popham (1988), for example, argues that a teacher could be doing all the right things, yet still get abysmal results in terms of student learning.

*Principle Two: Developmental Versus
Uniform Evaluation Criteria*

The second principle involves specifying multiple evaluation criteria designed to reflect a teacher's (or a group of teachers') stage of development, rather than formulating a single set of criteria and then uniformly applying it to all teachers. If we are to take the notion of teacher career development seriously, we must abandon the typical practice of using identical criteria to evaluate beginning and senior teachers. Through articulating evaluation criteria that reflect different development levels, districts may achieve greater clarity in communicating skill expectations to teachers and ensuring increased growth for all teachers.

Using the framework presented here, for example, at the application (or beginning) stage of a teacher's career, a criterion might be the "ability to respond to unanticipated classroom situations." A teacher at this level might be expected to deal well with interruptions and make an adjustment to an unanticipated situation. At the adaptation stage, the teacher could be expected to demonstrate greater flexibility. A criterion at this level might be the "ability to choose from a variety of alternatives in responding to unanticipated classroom situations." This criterion would measure the teacher's ability to integrate unexpected events into planning and use alternatives to take advantage of unanticipated situations¹. A leadership criterion might be the "ability to share classroom strategies with others." Thus, the criteria would be designed to reflect district expectations for teachers' skill development at different stages of their careers.

*Principle Three: Subjective
Versus Objective Evaluations*

The third principle involves acknowledging the subjective nature of teaching in the teacher-evaluation process. The source of such subjectivity should be readily acknowledged: it is difficult to achieve an objective system of evaluation in education because teaching is not simply the technical application of a set of clearly defined procedures for clear and predictable problems. Rather, teaching involves the exercise of judgment in the selection of alternative solutions in inherently uncertain situations.

A teacher's performance cannot be observed and evaluated without someone making judgments about the teacher's choices of alternative solutions. That person(s) may be the one who specifies the criteria and standards for an evaluation system; alternatively, it may be the person who actually performs the evaluation. Thus, the framework presented here magnifies the problem of subjectivity. Unlike most other evaluation schemes (that are designed to reflect some particular model of what constitutes effective teaching practices), this framework allows districts to design evaluation systems that emphasize the ambiguity and unpredictability inherent in teaching situations. Only when districts abandon the notion of "judgment-free" evaluation systems are evaluation systems likely to become more than simply sanctioning devices.

*Principle Four: Formative
Versus Summative Evaluations*

The fourth principle involves using the evaluation system as a set of diagnostic techniques designed to encourage teacher improvement, rather than as a process formulated to produce an overall pro or con assessment of a teacher. Two features of formative evaluation should be noted. First, formative evaluation involves more active participation by the teacher to be evaluated than does the more typical (summative) performance evaluations. Second, formative evaluation provides a way for teachers and evaluators to identify each teacher's strengths and weaknesses as a basis for further growth.

The framework presented here is well suited to formative evaluation, as the matrix format provides teachers and evaluators with a useful diagnostic tool to diagnose patterns of strengths and weaknesses in a teacher's decision-making skills. It is possible to determine, for example, that a teacher's strengths and weaknesses tend to concentrate in different rows or columns of the matrix. Thus, while systems based on competitive standards (discussed earlier) tend to use summative evaluation, systems based on developmental standards primarily focus on formative evaluation (Bacharach, Lipsky, & Shedd, 1984).

Implementing Developmental Standards

Although much of the reasoning and many of the details of the model presented here might be implemented without changing present systems of teacher compensation, one of its purposes is to provide an alternative to the so-called merit-pay and career-ladder plans presently being considered in several states. The proposal presented here assumes that these criteria would be employed in part to determine when a teacher is ready for promotion from one career level to the next, in a three-tiered system in which advancement is based on assessments of the level of a teacher's skills. Higher levels of compensation would be associated with each successive level.

It is important that there be no quotas on the number of teachers who might be promoted from one-career level to another, nor would there be any quotas on how many might be "at" one level or another at any one time. The standards for promotion should be demanding, but arbitrary quotas would be inconsistent with the basic purpose of the system, which is to encourage all teachers to develop their skills as much as possible. Obviously, the linkage between the stages of teacher career development and teacher compensation (and the further need to eliminate quotas) demands a continuing commitment to increases in teacher salaries commensurate with their skill development.

Furthermore, there would be no formal or direct relationship between the nature of a teacher's specific assignments or responsibilities and his or her career level. No conclusions could or should be drawn about a teacher's qualifications for particular assignments from the career level that he or she occupies. Career levels are meant to reflect a person's overall level of skills and development in the job of teacher. Using career level promotions as a way of breaking that one job into several more specialized jobs would negate the principle of continuous growth in a single profession and would generate a quota system.

On average, teachers at higher levels of the system might be more likely to have the specialized knowledge and leadership or innovation skills needed to perform some special duties, but a beginning teacher might have the special qualifications for a particular assignment. It is just as possible that a teacher might take on (or be asked to take on) a special assignment in order to develop such knowledge and skills. In any event, nothing in the structure of a career-development system should imply that special assignments (which often involve work outside a classroom) are in any sense more professional or require a higher level of skill than duties typically performed in the classroom. Formal restrictions on which assignments belong to which level of the system, therefore, would be inconsistent with the system's basic structure and purpose.

Conclusion

We began this chapter by maintaining that many of the changes in teacher compensation have failed to take into account the purpose of the evaluation system. In contrast to past approaches emphasizing minimum standards, this chapter argues that an evaluation system must do more than motivate teachers to perform at the least-common-denominator level. At the same time, an evaluation system should not set teachers against each other for scarce rewards. In contrast to minimum and competitive standards approaches, this chapter is based on the assumption that an evaluation system in education should attempt to develop all teachers to their maximum capacity.

What does it mean to develop teachers? We feel that successful teachers are those who can plan, implement, and evaluate decisions in performing three critical roles — instruction, management and counseling. We further argue that evaluation systems for development should be skill based rather than perfor-

mance based, should maintain developmental rather than uniform evaluative criteria, should recognize the subjective component in evaluation systems, and should place primary emphasis on formative evaluations instead of summative evaluations.

For such an evaluation system to be successful, what ultimately has to change is the organizational culture of the school. Any evaluation system directed at teacher development will fail unless that system is not embedded in an organizational culture where management is committed to its professional staff. The bureaucratic culture and conflictive atmosphere that currently dominate our schools inevitably breed distrust. Without trust, any evaluation system, no matter how well intended, is doomed.

Notes

1. Although the career stages in this paper are loosely based on those proposed by Burden (1982), other observers trace career development patterns in teaching in ways that are analogous. (For examples, see Fuller, 1970; Fuller & Brown, 1975; Newman, 1979; and Peterson, 1978).
2. The principles are not mutually exclusive, indeed they overlap substantially; for example, developmental criteria are generally also formative and skill-based in nature.
3. This skill involves adaptation because it requires teachers to have a variety of techniques available to them (i.e., the first feature of the ability to adapt); it also requires teachers to integrate planning, implementation and evaluation decisions (i.e., the second feature of the ability to adapt).

References

- Bacharach, S. B., & Conley, S. C. (1988). Uncertainty and decisionmaking in teaching: implications for managing line professionals. In T. Sergiovanni & J. H. (Eds.) *Schooling for tomorrow: Directing reforms to issues that count*. 311-328.
- Bacharach, S. B., Conley, S. C., & Shedd, J. B. (1986). Beyond career ladders: Structuring teacher career development systems. *Teachers College Record*, (Summer) 87(4), 563-574.
- Bacharach, S. B., Conley, S. C., & Shedd, J. B. (1987). A career developmental framework for evaluation of teachers as decision makers. *Journal of Personnel Evaluation in Education*, 1(1), 181-194.
- Bacharach, S. B., Lipsky, D. B., & Shedd, J. B. (1984). *Paving for better teaching: Merit pay and its alternatives*. Ithaca, NY: Organizational Analysis and Practice.
- Bacharach, S. B., Shedd, J. B., Malanowski, R., & Conley, S. C. (1984). *Criteria and standards for a teacher career development system*. Report prepared for the National Education Association. Ithaca, NY: Organizational Analysis and Practice.
- Burden, P. R. (1982). Developmental supervision: Reducing teacher stress at different career stages. Paper presented at the annual meeting of the Association of Teacher Educators, Phoenix, Arizona. ERIC ED 218.
- Calhoun, F. S., & Protheroe, N. J. (1983). *Merit pay plans for teachers: status and descriptions*. Arlington, VA: Educational Research Service.
- Christenson, J., Burke, P., Fessler, R., & Hagstrom, D. (1983). Stages of teachers careers: Implications for professional development. National Institute of Education, Washington, DC. (monograph). ERIC SP 021 495.
- Darling-Hammond, L. (1986). A proposal for evaluation in the teaching profession. *The Elementary School Journal*, 86(4), 1986, 531-551.
- Doyle, W. (1985, April). *Teaching as a profession: What we know and what we need to know about teaching*. Paper presented at the annual meeting of the American Educational Research Association, April, Chicago.

- Educational Research Service, Inc. (1979). *Merit pay for teachers*. Arlington, VA: Author.
- Fuller, F. (1970). *Personalized education for teachers: An introduction for teacher educators*. University of Texas, Austin, Research & Development Center for Teacher Education. Office of Education (DHEW), Washington, DC. ERIC ED 048 105.
- Fuller, F. & Brown, O. (1975). On becoming a teacher. In K. Ryan, (Ed.) *Teacher Education*, 74th Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.
- Gage, N. L. (1972). An analytic approach to research on instructional methods. In A. Morrison & D. McIntyre (Eds.) *The social psychology of teaching*. Hammondsworth, Eng.: Penguin Books.
- Glassberg, S. (1979). A developmental model for the beginning teacher. In K. R. Howey & Richard H. Bents (Eds.) *Toward meeting the needs of the beginning teacher*. Washington, DC: Office of Education, Midwest Teacher Corps, Eric ED 206 581, 111-138.
- Hackman, J. R. & Oldham, G. (1980). *Work redesign*. Reading, MA: Addison-Wesley.
- Hall, D. T. (1976). *Careers in organizations*. Pacific Palisades, CA: Goodyear Publishing Co.
- Hart, A. (1987). A career ladder's effect on teacher career and work attitudes, *American Educational Research Journal*, 24(4), 479-503.
- Hatrey, H. & Greiner, J. M. (1984). *Issues in teacher incentive plans*. Washington, D.C.: The Urban Institute.
- Hunter, M. (1988). Create rather than await your fate in teacher evaluation. In S. Stanley and W. Popham, (Eds.) *Teacher evaluation: Six prescriptions for success*, 32-35. Alexandria, VA: Association for Supervision and Curriculum Development.
- Iwanicki, F. (1981). A Professional-growth oriented approach to evaluating teacher performance, J. Millman, (Ed.) *The handbook of teacher evaluation*, 203-228. Beverly Hills, CA: Sage Publications.
- Jackson, P. (1968). *Life in classrooms*. New York: Holt, Rinehart & Winston.
- Lampert, M. (1985, May). How do teachers manage to teach: Perspectives on problems in practice. *Harvard Education Review*, 55, 178-194.
- Lortie, D. (1975). *School teacher: A sociological study*. Chicago: University of Chicago Press.
- McPherson, G. (1972). *Small town teachers*. Cambridge: Harvard University Press.
- Malen, B. & Hart, A. (1987). Career ladder reform: a multilevel analysis of initial efforts, *Educational Evaluation and Policy Analysis*, 9, (1), 9-23.
- Metz, M. H. (1978). *Classroom and corridors: The crisis of authority in desegregated secondary schools*. Berkeley, CA: University of California Press.
- Murnane, R. J. & Cohen, D. K. (1986). Merit pay and the evaluation problem: why some merit plans fail and a few survive. *Harvard Educational Review*, 56, 1-17.
- Newman, K. K. (1979). *Middle-aged experienced teachers' perceptions of their career development*. Unpublished doctoral dissertation. Columbus, OH: Ohio State University.
- Peters, T. & Waterman, R. H. (1982). *In search of excellence*. New York: Harper and Row.
- Peterson, A. R. (1978). *Career patterns of secondary school teachers: An exploratory interview study of retired teachers*. Unpublished doctoral dissertation. Columbus, OH: Ohio State University.
- Popham, W. J. (1988). Judgement-based teacher evaluation. In S. Stanley and W. Popham, (Eds.) *Teacher evaluation: Six prescriptions for success*, 62-63. Alexandria, VA: Association for Supervision and Curricular Development.
- Rosenholtz, S. J. & Smylie, M. A. (1984). Teacher compensation and career ladders. *Elementary School Journal*, 85, 149-166.
- Schlechty, P. (1989). Career ladders: a good idea going awry. In T. J. Sergiovanni and J. H. Moore (Eds.) *Schooling for tomorrow*. Boston: Allyn & Bacon.
- Slavin, Robert E. (1984). Component building: A strategy for research-based instructional improvement. *Elementary School Journal*, 84, 255-269.
- Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). Teacher evaluation: A study of effective practices. Santa Monica, CA: The RAND Corporation.
- Zeichner, K. M. (1983). Alternative paradigms of teacher education. *Journal of Teacher Education*, 34, 3-9.
- Zumwalt, K. (1982). Research on teaching: Policy implications for teacher education. In A. Lieberman and M. McLaughlin (Eds.) *Policy making in education*, 81st Yearbook of the National Society for the Study of Education, Part 1, Chicago: University of Chicago Press.

9. Evaluation for Tenure and Dismissal

Edwin M. Bridges

The context for teacher evaluation in the 1990s is now taking shape. Large numbers of teachers will be hired in a relatively short period of time during a period of teacher scarcity. There is the ever present danger that the events that transpired in the 1950s, 1960s, and 1970s will be repeated in the 1990s. If this pattern recurs, these newly hired teachers will not be closely supervised during their probationary period and will acquire tenure even if there are serious doubts about the adequacy of their performance in the classroom. Once these teachers have obtained tenure, their evaluation will be driven primarily by parent and student complaints. The students who are most likely to be shortchanged by this eventuality will be the students who need a quality education the most — minorities and students from low socioeconomic backgrounds. To prevent this unwelcome possibility from becoming a reality, local school districts must replace the complaint-driven approaches of the past with systems of evaluation that reflect the importance of merit and equity in relation to tenure and dismissal decisions.

Evaluation: Past Practices

During the period of the last teacher shortage, most school districts had relatively weak formal systems of teacher evaluation (Bridges, 1986).¹ In these systems there were no clear, explicit criteria for evaluating teachers; principals were not adequately prepared to handle their evaluative function, and they were not held accountable for how well they were performing this key function. Lenient evaluation practices resulted, and marginal teachers acquired tenure with little or no thought given by the administration to the long-range consequences of this action. Administrators, if they had doubts about the wisdom of their decisions, rationalized these reservations away on the grounds that "the

AUTHOR'S NOTE: I wish to acknowledge and assistance of Dale Bolton, Barry Groves, and Susan Johnson who graciously provided valuable comments on earlier drafts of this manuscript.

next one could even be worse" or "all this teacher needs is a little more seasoning."

These teachers, once ensconced in the legal cocoon of tenure, became the unwitting subjects of a complaint-driven approach to teacher evaluation. This nonformal approach evolved as an administrative response to the problems inherent in evaluating and dealing with teachers who were entitled to due process in matters of employment. In a complaint-driven approach teachers — irrespective of their performance in the classroom — received favorable feedback and ratings from principals until the teacher became the target of numerous complaints from parents and students. When these complaints surfaced, the teacher's subsequent evaluations and treatment depended to a large extent on the availability of escape hatches, that is, ways to avoid current and future complaints. If an escape hatch were available, the teacher's poor performance was tolerated; if the escape hatches were closed, the teacher was confronted about his or her deficiencies in the classroom. Let us now take a more detailed look at how the elements of this complaint-driven approach worked in actual practice.

In a complaint-driven approach to teacher evaluation, principals were inclined to tolerate the inadequacies of teachers unless there were complaints, especially from parents. This tolerance manifested itself in one of several ways. Principals ignored or overlooked the poor performance, filed written observation reports sprinkled with glowing generalities such as "I really enjoyed my visit," used double-talk in written evaluations to muffle criticism of the teacher's performance, and inflated performance ratings in the mistaken belief that these evaluations would act as positive reinforcement. By dealing with marginal performance in these ways, principals spared themselves the unpleasantness involved in communicating negative feedback and avoided the need to spend countless hours with the teacher in an effort to improve his or her performance. Given the principal's broad range of responsibilities and large span-of-control (i.e., number of employees to supervise), these tolerant responses were understandable, even if not acceptable.

If parental complaints arose and the principal judged them to be legitimate, the administration generally preferred to take the path of least resistance and to use escape hatches rather than confront the teacher. Transferring the teacher to another school was a favorite escape hatch. Its popularity stemmed from the multiple purposes it served. The practice took the heat off the immediate supervisor, spared the administration from having to confront the poor performer, and protected the district against defenses commonly used by teachers in dismissal proceedings, namely, "My supervisor had it in for me" or "My supervisor and I had different philosophies." Administrators referred to this practice as "the dance of the lemons."

When trying to find a new home for the teacher who was under fire, administrators looked for schools where the parents were not likely to complain. Schools that met this criterion were ones with high rates of transiency and large concentrations of low SES students and disadvantaged students, usually from Black and

Hispanic populations. Middle-class sites were avoided because parents of children in these schools were far more likely to voice their complaints and press for action if the complaints were not handled to the parents' satisfaction.

If a "suitable" site were not available, administrators resorted to other escape hatches, work assignments that were unlikely to be a breeding ground for parent and student complaints. For example, the marginal teacher was assigned to teach electives; in this way potential complainants could avoid the teacher if they so chose. Alternately, the teacher was assigned to work as a roving substitute. By using weak teachers in this capacity, districts solved two problems at once. They filled their daily needs for substitutes, who were in short supply, and avoided complaints by shifting the teacher from one classroom to another on a daily basis. Ineffective teachers also served as home-study teachers and worked one-on-one with students in their homes. Parents of these children were grateful, not critical, of what was happening. Finally, ineffective teachers were reassigned to nonteaching positions such as working in the office on a special project or in a noncertificated role (e.g., secretary, custodian, or school bus driver). In all of these positions, teachers were less likely to engender parental dissatisfaction and complaints.

These various escape hatches were not always available, however. Their availability depended upon the size and fiscal health of the district. Strapped for funds, districts could not afford to assign marginal teachers to tutor students in their homes, to teach electives with small enrollments, to work on special projects in the central office, and to work in noncertificated roles as extra help. Small districts, unlike large ones, were unable to transfer ineffective teachers between schools and to use teachers in the roving substitute pool.

If the escape hatches were closed because of one or both of these conditions, administrators confronted teachers when complaints arose. These confrontational responses fell into two distinct stages: (1) how to improve the teacher's performance, and (2) how to get rid of the teacher if she or he failed to improve. Teachers who showed few signs of improvement were generally not dismissed; rather, they were induced to resign or to retire early.

The remedial attempts produced little, if any, improvement; their limited success was attributable to a number of factors. These efforts were usually accompanied by unmutated criticism from the principal and detailed specifications of what and how the teacher should fulfill his or her responsibilities in the classroom. Having been fed heavy doses of glowing generalities, double talk, and inflated ratings in the past, these teachers became defensive and resisted the efforts to help them improve. The principals, hampered by the lack of organizational resources and an adequate technology for diagnosing and remediating the poor performance, were unable to cope effectively with these defensive reactions or to provide adequate assistance. Under these conditions, it is hardly surprising that teachers rarely improved and, even when they did, the improvement was not dramatic.

When these improvement efforts failed to produce the desired results, administrators attempted to get rid of the teachers. Administrators preferred induced resignations and early retirements to dismissal. These induced exits were usually preceded by a period of intense pressure, negotiations to determine the terms and conditions of the teacher's departure, cooperation behind the scenes between the union and the administration, and inducements to the teacher in return for the resignation or the early retirement. The most common inducements were extended medical coverage, employment as a consultant for a fixed period (usually five years), cash, and subsequent employment as a substitute teacher.

A relatively small percentage of the teaching force — less than 1% a year — was forced to resign or to retire early under this complaint-driven approach. As a result, nearly two million students a year were being shortchanged by marginal performance in the classroom, enough students to fill all of the classrooms in the 14 smallest states.² Moreover, the students who were most likely to be cheated by the escape hatches inherent in this approach were minorities and students from low SES backgrounds.

Evaluation for Tenure: An Approach Based on Merit and Equity

As school districts enter the next decade, it is imperative that they learn from and act on the lessons of the past. The period from which school districts have recently emerged underscores the cruciality of the tenure decision. It is the most important personnel decision a district makes because of its profound impact on the quality of the teaching force, the district's flexibility in responding to changing circumstances in the future, and the equality of educational opportunities.

The tenure decision represents the last occasion for a district to impose a reasonable standard of performance on its teachers. Once a teacher has acquired tenure, the courts and the hearing officers presume that a teacher is competent. Incompetence, not competence, becomes the standard by which the teacher's continued employment is judged. Unless the district is able to prove that the teacher is incompetent, the teacher is legally entitled to continue teaching. Incompetence does not mean marginal or mediocre; rather, it means that the person is a blatant failure in the classroom — so poor that no reasonable person would doubt that the teacher is inept in the classroom. Until the standard of performance for revoking tenure is raised by state legislatures,³ school districts that are concerned about the quality of their teaching force necessarily must look to the tenure decision as the major leverage point.

Tenure decisions are also important because they limit the future flexibility of districts in responding to the changing realities of organizational life. Enrollments rise and fall, fiscal resources fluctuate, societal needs and expectations change, and the character of student populations alters, sometimes dramatically.

Each decision to grant tenure constrains the ability of a district to deal intelligently and responsibly with these organizational vicissitudes.

Finally, tenure decisions are important because the negative consequences of poor decisions are likely to be borne by the students who are the most vulnerable in the system. These students are the ones who are most often victimized by the complaint-driven approach to teacher evaluation — the poor, the minorities, and the low achieving.

In designing a system of teacher evaluation that reflects the crucial importance of the tenure decision, there are at least two ways of approaching this problem. One is to ask the question, "Are there grounds for denying the person tenure?" This way of formulating the problem is essentially how the decision to dismiss the teacher, once she or he has been granted tenure, is framed. This negative approach to decision directs evaluators to the weaknesses in the teacher's performance and ensures that weak teachers, if their shortcomings are identified and documented, will not be granted tenure.

An alternate way of framing the problem is to pose the question as follows, "Is there compelling evidence that the person deserves tenure, i.e., is capable of creating a satisfactory learning environment for all types of students — poor children as well as affluent children, minority students as well as white students, low-achieving as well as high-achieving students? This approach to the problem focuses on establishing whether there is sufficient evidence to induce a reasonable belief that the person is competent in carrying out his or her role as a classroom teacher with different types of students. It directs evaluators to search for evaluation of effectiveness, not ineffectiveness; moreover, it underscores the value of proving one's competence in multiple contexts — contexts differentiated by the type of students being taught.

The second approach is far more likely to accomplish the goals of a good-quality teaching force and equality of educational opportunities for students. This approach, unlike the first one, challenges two major assumptions underlying current practices. First, it does not assume that the absence of documented deficiencies is proof of one's competence in the classroom. Second, it does not assume that evidence of a person's competence in working with one type of student proves that one is able to work effectively with all types of students. By questioning these suspect assumptions and demanding compelling evaluation to substantiate a recommendation for tenure, the second approach increases the chances of granting tenure to teachers capable of providing a good-quality education to all students.

In order to implement this merit- and equity-based approach to the tenure decision, local school districts necessarily must grant tenure only when the following three conditions are met:

1. Multiple sources of evidence (e.g., supervisor observations, student ratings, student test performance, and observations by peers who are expert in the context in which the teacher is being observed) are used to evaluate the teacher's competence

in the classroom, and these multiple sources point to the same conclusion, namely, the person is a fully competent teacher.

2. Each source of evidence is obtained at several different points in time, and the repeated measures of performance yield similar results with respect to the person's competence.
3. The teacher is evaluated in multiple contexts, contexts distinguished by types of students being taught (e.g., social class, ethnicity, and ability), and the teacher demonstrates his or her competence in these different contexts.

Given these stringent standards for judging a teacher's suitability for tenure, it is imperative that teachers be evaluated under conditions conducive to fully competent performance. The teacher's assignment should represent a reasonable level of difficulty with respect to the number of preparations, the number of students who are viewed as difficult to teach, and the number of locations in which they are expected to teach. Moreover, the performance should be judged only in classes within the teacher's area of certification. Finally, the teacher should have adequate resources (e.g., equipment, instructional materials, and facilities) to carry out his or her teaching responsibilities. If these conditions are not met, it is unfair to hold the teacher responsible for a level of performance that is not possible under the circumstances.

Evaluation Beyond Tenure: Dismissal

Dismissal of a tenured teacher for poor performance is a relatively rare event (Bridges, 1986); it should be even rarer if the district has adopted a merit- and equity-based approach to the tenure decision. Despite the rarity of dismissal, school districts must consciously design their evaluation system with this possibility in mind. This design should emphasize practices which do the following: (1) enhance the opportunities for the poor performer to improve and, thereby, reduce the exposure of the teacher and the district to the costs and pain inherent in dismissal; (2) heighten the probability that a dismissal decision will be sustained if it is contested; and (3) increase the likelihood that minorities and students from low socioeconomic backgrounds will receive equal educational opportunities.

Effective Problem-Solving Process

Solving serious performance problems usually goes through seven distinct stages. In the first stage, the poor performer *ignores* the problem until facts, examples, and evidence are introduced to demonstrate that a problem exists. When confronted with this evidence, the teacher *denies* that the problem exists and challenges the accuracy of the facts or indicates that they grossly exaggerate what has happened. The period of denial is often followed by *attacks* on the

supervisor and accusations that the supervisor is unreasonable, unprofessional, and out to get the employee. Once the attacks have subsided, the employee usually acknowledges that a problem exists; however, she or he *blames others* for what has occurred. If the problem-solving process is not aborted at this stage, the poor performer eventually *assumes responsibility* for the problem and acknowledges that the fault lies with him or her, not others. Having assumed full or partial responsibility for the problem, the poor performer, along with the supervisor, *finds a solution* to it. In the seventh and final stage of the process, the teacher *implements* the solution to the progress; the process terminates if the solution proves successful.

This problem-solving process commonly breaks down at or prior to stage 4 (blaming others) because the supervisor does not understand the usual progression of the process involved in dealing with serious performance problems (Grove, 1985). It is important for supervisors to recognize where the teacher is in this process, to avoid jumping to the solution stage before the teacher has assumed responsibility for the problem, to resist the temptation to respond in kind when the teacher is being defensive (i.e., denying, attacking, or blaming others), and to facilitate the transition of the teacher from one stage to the next by listening and helping the teacher to recognize the typicality of his or her responses and the consequences of failing to face up to the problem.

Appropriate Assistance

Tenured teachers who experience major difficulties in the classroom and are provided with remedial assistance rarely show much improvement (Bridges, 1986). The lack of improvement is attributable in part to the nature of the causes for their poor performance and the inappropriate nature of the assistance they receive. Given the lack of an adequate technology for dealing with ineffective performance, it is understandable why these rescue attempts are often misguided.

Teachers who are having serious problems in the classroom commonly receive assistance that is rooted in the assumption that the teacher's problems stem from individual knowledge deficiencies and skill deficiencies. For example, the teacher may be given written specifications about how to behave in the classroom in relation to teaching tasks and methods, provided with opportunities to visit the classes of exceptional teachers, released from regular classroom duties in order to attend workshops, and granted access to consultants for a short period of time. Although this type of assistance is legally defensible, it is unlikely to be effective unless the teacher's difficulties stem solely from knowledge deficiencies and/or skill deficiencies.

The origins of ineffectiveness are often complex, not simple. Seldom is the source of difficulties in the classroom attributable to a single cause such as lack of effort, skill, or ability. Rather, the poor performance has multiple causes. Besides insufficient ability or motivation, poor performing teachers often suffer

from some type of personal disorder or pathology (e.g., emotional stress, burn-out, health problems, or alcoholism) that adversely affects their performance. Outside influences such as marital difficulties, problems with children, and financial woes also contribute to the teacher's difficulties in the classroom. Finally, improper teaching assignments and working conditions, combined with inadequate supervision, may be undermining the teacher's classroom performance. Because the teacher's problems commonly stem from several factors, remedial assistance, if it is to be effective, must be targeted toward all of the underlying causes, not simply the presumed lack of skill or ability. Evaluations necessarily should seek to pinpoint the causes for this poor performance and to prescribe assistance that reflects this diagnosis.

Credible Evidence

If the teacher's difficulties persist despite good faith attempts by the district to manage the problem-solving process effectively and to prescribe appropriate assistance, the administration must be able to justify its dismissal decision to an impartial third party (a court judge, a hearing officer, or a commission on professional competence). A key ingredient of a legally defensible dismissal decision is credible evidence. To justify its decision, the administration must be able to establish through its written records that a pattern of poor performance exists in relation to the district's criteria for evaluating teachers. Because there are no clearcut standards or yardsticks for determining whether a teacher is meeting a particular criterion, administrators must accumulate numerous examples of the teacher's shortcomings and use these specific instances to verify that there is a pattern of unsatisfactory performance (Bridges & Groves, 1984). The significance of a demonstrable pattern is underscored in the following statement by an appellate court judge:

Proof of momentary lapses in discipline or of a single day's lesson gone awry is not sufficient to show cause for dismissal of a tenured teacher. . . . Yet, where brief instances and isolated lapses occur repeatedly, there emerges a pattern of behavior which, if deficient, will support the dismissal of a tenured teacher. Where the school board fails . . . to show that the examples of conduct constitute a pattern of deficiency, then dismissal cannot be permitted (Board of Education v. Ingels, 394 N. E. 2d 69, 1979).

In addition to demonstrating that a written record of recurring deficiencies exists, the administration needs written proof to verify the following (Frels & Cooper, 1982; Carey, 1981):

1. the teacher received copies of the relevant documentation;
2. the documentation was delivered in timely manner;

3. the teacher was given an opportunity to refute or comment on what the supervisor has written;
4. the persons who have filed written complaints will later testify to their authenticity.

The credibility of the written record also will be enhanced if the classroom observations are conducted by more than one supervisor, one of whom is expert in the teaching context in which the teacher is being observed. Other sources such as parental and student complaints, student test results, lesson plans, and student ratings may be used to provide the necessary background for understanding the performance deficiencies of the teacher; however, the legal standing of these sources does not appear to equal the evaluations of supervisors that are based on classroom observations (Bridges & Groves, 1984).

Procedural Due Process

Prior to being dismissed, a tenured teacher is entitled to procedural due process. These procedural rights are generally spelled out in the state education code. Although the specific details vary from one state to another, the major elements remain virtually the same. A tenured teacher is entitled to a fair hearing, which entails a timely notice of reasons and charges, representation by legal counsel, an opportunity to cross-examine witnesses, an impartial hearing by individuals who do not present or prosecute the case for dismissal, and a decision that is based only on evidence presented at the hearing. When districts fail to provide teachers with these due process rights, the courts or hearing officers are likely to overturn the dismissal decision.

Adequate notice is the cornerstone of due process. Teachers have the right to know in advance what the district's rules and expectations are for a teacher's conduct and performance. Moreover, teachers need to be informed about the specific ways in which their conduct is deficient. Merely stating that a teacher's performance should be improved is not enough.

Without knowledge of the specifics in which classroom conduct is deficient, a teacher who seeks to improve his or her teaching ability may find that such efforts result in classroom conduct that in the minds of school authorities, is even less competent, less efficient . . . There is no assurance that any particular course of action undertaken by a teacher . . . will constitute sufficient improvement in the eyes of the board and school authorities. The teacher finds herself struggling blindly towards undefined and unknown standards of conduct (Pollard v. Board of Edu. Reorganized School District, 533 S.W., 2d 667, 1976).

Restrictive Transfer Policies

The Fourteenth Amendment to the U. S. Constitution contains two clauses, one having to do with due process and the other with equal protection. This important civil rights amendment was passed after the Civil War to safeguard the interests of the newly freed slaves. Ironically, this amendment on occasion has resulted in procedural due process for tenured teachers in matters of employment and unequal educational opportunities for educationally and socially disadvantaged students. The complaint-driven approach to evaluation, as noted earlier, has resulted in the assignment of ineffective teachers to classes and schools where parents are unlikely to complain. These generally are classes and schools attended by students from minority and low socioeconomic backgrounds.

If the promise of the Fourteenth Amendment is to be fulfilled for these students and the large gaps in educational achievement and attainment between children of different racial, ethnic, and socioeconomic backgrounds are to be closed, school districts must adopt transfer and reassignment policies that protect the rights of the disadvantaged to an equal education. To counter the practice of using schools for the disadvantaged as sites for the annual "dance of the lemons," districts should adopt transfer policies such as the following: if a teacher receives a positive evaluation in one school, but nonetheless is transferred to another, and then encounters difficulty in the new setting, the teacher should be returned to the first school, and that school's principal should be placed under surveillance. Seemingly harsh policies such as this one should prevent schools serving disadvantaged students from becoming a dumping ground for the teachers who are having difficulties in more advantaged school settings. Moreover, closing these escape hatches should stimulate all principals to deal more forthrightly with the poor performers.

Conclusion

If the evaluation practices recommended here in relation to tenure and dismissal are to be implemented, districts must create several enabling conditions (McLaughlin & Pfeifer, 1988). Evaluators will need to be trained: to carry out their responsibilities; to be held accountable for making these necessary, but at times unpleasant, decisions; to have access to the resources (e.g., time, legal counsel, and authority) essential to fulfilling their evaluative functions effectively; and to be assured of support from the superintendent and board of education if decisions encounter opposition from vocal members of the school community. Unless these enabling conditions are created, the complaint-driven approach to evaluation that has operated in most districts during the past three decades will continue. If this happens, districts unwittingly will have slammed shut the window of opportunity afforded by the upcoming influx of new teachers into the profession. The nation's students, especially the minorities and the disadvantaged, deserve a better fate.

Notes

1. This entire section is based on studies reported in Bridges (1986).
2. The number of students was estimated as follows. In the fall of 1981 there were 2,124,697 full-time equivalent teachers employed in the United States (Grant & Snyder, 1984). Five percent of this figure (a conservative estimate of the percentage of teachers who are incompetent) is 106,235, the estimated number of incompetent teachers. I assumed that each of these teachers taught 18.9 students (the average pupil-teacher ratio reported for 1981-82). Under this assumption, the total number of students being taught by incompetent teachers is 2,007,842. This number exceeds the total combined public school enrollments reported in 1981-82 for the 14 smallest states.
3. Legislatures should adopt a higher standard of performance, and this standard is marginal performance, not incompetence. (Marginal performance means that the teacher's performance falls just short of fulfilling one or more of the professional duties of a teacher.)

References

- Bridges, E. M. (1986). *The incompetent teacher*. Philadelphia: Falmer Press.
- Bridges, E. M. & Groves, B. (1984). *Managing the incompetent teacher*. Eugene, Oregon: ERIC Clearinghouse on Educational Management.
- Carey, W. (1981). *Documenting teacher dismissal*. Salem, Oregon: Options Press.
- Frels, K. & Cooper, T. (1982). *A documentation system for teacher improvement or termination*. Topeka, Kansas: National Organization on Legal Problems of Education.
- Grant, W. & Snyder, T. (1984). *Digest of education statistics 1983-1984*. Washington, DC: National Center for Educational Statistics.
- Grove, A. (1985). *High output management*. New York: Random House.
- McLaughlin, M. & Pfeifer, R. (1988). *Teacher evaluation: Improvement, accountability, and effective learning*. New York: Teachers College Press.

10. Teacher Evaluation for School Improvement

Edward F. Iwanicki

Purposes Are the Foundation of the Teacher-Evaluation Process

There are three basic components of an effective teacher-evaluation process: (1) philosophy and purposes; (2) performance criteria and standards; and (3) evaluation procedures. The philosophy and purposes convey the role and function of teacher evaluation in the educational decision-making process. They state why teachers are evaluated. Performance criteria and standards specify what is expected of teachers in their professional roles. Evaluation procedures describe how teachers are assessed. The purposes for teacher evaluation have a direct effect on the performance expectations we set for teachers as well as on how these professional staff are evaluated. Thus, the purposes are the foundation of the teacher-evaluation process. Any school strengthening its teacher-evaluation process should begin by affirming its purposes for teacher evaluation. Once clear purposes are affirmed, performance expectation for teachers and evaluation procedures should be reviewed to determine whether they are congruent with these purposes. Any mismatches between performance expectations or evaluation procedures and the purposes of teacher evaluation should be resolved.

AUTHOR'S NOTE: The author expresses his thanks and appreciation to the following colleagues for the constructive feedback they provided during the preparation of this manuscript: David L. Clark, University of Virginia; Thomas L. McGreal, University of Illinois; Christine Murphy, University of Connecticut; Philip A. Streifer, Barrington (RI) Public Schools; Robert M. Villanova, Farmington (CT) Public Schools. Also, it is important to acknowledge the contributions numerous other colleagues have made to the development of the views presented in this chapter.

The preparation of this manuscript was supported in part by funds from the United States Department of Education under the Leadership in Educational Administration Development (LEAD) act of 1984 through the Connecticut LEAD Center at ACES. The opinions expressed here do not necessarily reflect the positions or policies of USDE, and no official endorsement should be inferred.

The author expresses his thanks and appreciation to the many school system personnel who have had input into the development of the Teacher-Evaluation Cycle. A version of this cycle is being implemented or being considered for implementation in the following Connecticut school systems: The Connecticut Vocational-Technical Schools, Enfield Public Schools, Norwalk Public Schools, Plymouth Public Schools.

This perspective that the philosophy and purposes are the foundation of the teacher-evaluation process is critical. Too often school systems become caught up in activities such as revising position descriptions, improving observation protocols, or simplifying evaluation forms, activities more related to how teachers are evaluated — when the real problem is a lack of common understanding between teachers and administrators as to the real purposes of the teacher-evaluation process. A unified vision of why teachers in a school or district are being evaluated is necessary before efforts to improve evaluation procedures can have much impact. Often teachers' resistance to evaluation is not motivated by the particular performance criteria or procedures being applied, but rather, by the purposes for which the data gathered through the teacher-evaluation process can be used. There must be agreement on these purposes before an effective teacher-evaluation process can be implemented (Bolton, 1973; McGreal, 1983; Stiggins & Bridgeford, 1985; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984).

As seen in this and other works, there are numerous ways to state the purposes of teacher evaluation. Although lists of purposes vary in content and length, there are essentially four purposes for which teacher evaluation can be used. These are as follows:

1. Accountability — to ensure that only effective teachers continue in the classroom
2. Professional Growth — to foster the professional growth of new and continuing teachers
3. School Improvement — to promote school improvement and the enhancement of student learning
4. Selection — to ensure that the best qualified teachers are hired

Educators involved in the development of teacher-evaluation programs tend to place more emphasis on the accountability and professional growth purposes, and less emphasis on the school improvement and teacher-selection purposes. The optimal scenario with respect to these purposes would be to select high quality teachers and then induct these staff using an evaluation process that fosters professional growth and promotes school improvement. In settings where the teacher-evaluation process has a demonstrable impact on teacher selection, professional growth, and school improvement practices, less attention needs to be paid to the traditional accountability purpose of evaluation.

School Improvement As a Purpose of Teacher Evaluation

The focus of this chapter is on school improvement as a purpose of teacher evaluation. Just as the professional growth efforts of staff are often integrated

with the teacher-evaluation process, so too should school-improvement efforts be integrated with the teacher-evaluation process. Some educators may not accept school improvement as an appropriate purpose for teacher evaluation. They would argue that there are more effective ways to bring about school improvement than through the teacher-evaluation process. This may be the case, but the teacher-evaluation process can complement the particular change strategies used by a school system to bring about school improvement. For example, the literature on school effectiveness and school improvement (Loucks-Horsley & Hergert, 1985; Purkey & Smith, 1983) encourages teachers and administrators to become involved in planning and processes that are compatible with, and can be complementary to, those we pursue in effective teacher evaluation for professional growth. As one administrator commented after a conference presentation, "Putting the two processes together makes so much sense. The only thing that bothers me is why didn't we think of doing this ourselves. It's so logical." One reason why some administrators may not have thought about integrating school effectiveness or school improvement efforts with the teacher evaluation process for professional growth is their tendency to add new initiatives, rather than to examine how school-improvement efforts might fit in with the existing teacher-evaluation process.

It is especially appropriate to integrate teacher evaluation and school improvement in school systems where goal-setting models of teacher evaluation (Iwanicki, 1981; McGreal, 1983; Redfern, 1980) are being used. In such settings, teachers can be asked to make school improvement part of their professional goal structure and therefore, part of the teacher-evaluation process. School improvement provides staff with another option when writing their performance objectives. In many cases teachers have had only two options. They could state objectives to either (1) strengthen their performance with respect to the criteria included in their position description or (2) enhance their performance through the professional growth process. Experienced, successful teachers, who did not need to strengthen their performance with respect to their position description, did not have any options. They would have to state objectives to enhance their performance through professional growth. After writing growth-oriented objectives for two to three years, such teachers began to ask for other options. It was not that they chose not to grow further professionally. They would continue to pursue appropriate professional growth activities, but they were looking for another option when exerting that extra effort to improve the student learning experience. School improvement is a logical and attractive complement to professional growth for these teachers.

It is not easy to classify an objective as a professional-growth objective or a school-improvement objective by just looking at the objective, especially if it is a well-stated performance objective. For example, the following objective was developed by a seventh-grade language arts teacher:

I will complete the Connecticut Institute for Teaching and Learning workshop on process writing, and then apply the skills I have acquired to improve writing instruction in my classroom.

Depending on the circumstances, this objective could have both a professional-growth dimension (complete the workshop on process writing) and a school-improvement dimension (improve writing instruction in my classroom). This objective would be classified as a professional-growth objective if the teacher developed this objective in response to a particular need he had in his teaching. If a school priority had been set to address the teaching of writing and this teacher was completing the workshop to improve the writing curriculum in the school, it would be viewed as a school-improvement objective. The distinction being made is that performance objectives developed in response to individual teacher needs are classified as professional-growth objectives, while those objectives that address a school priority are viewed as school-improvement objectives.

Setting a Focus for School Improvement

When school improvement is a purpose of teacher evaluation, a focus must be set for the school-improvement process so that teachers can address improvement needs in a coordinated manner. This focus may be set by involving staff in a system-wide educational planning process (Kaufman, 1988; McCune, 1986). This process should result in building-level school-improvement goals, which are addressed over the next three-to-five years. For each goal, an action plan consisting of objectives and activities is developed by building staff, and procedures are specified for monitoring/evaluating progress toward these objectives. In many cases, the objectives are prioritized by building staff. The priority target objectives are those that address the most critical educational needs in that school. Furthermore, these target objectives tend to be ones where teachers want to invest their energies and resources in the future. Usually, no more than two, or possible three, target objectives are identified for the building.

There is no question that the educational needs of the students should take precedence when identifying priority objectives at building level. In situations where a large number of priority objectives have been identified on the basis of student needs, staff preferences could be used to reduce that number of objectives still further. Staff preferences are an important consideration when identifying target objectives. Teachers's efforts tend to have a greater impact when they pursue an objective "so stated that it excites the imagination and gives people something they want to work for, something they don't yet know how to do, something may can be proud of when they achieve it" (Kappel, 1960, p. 38). In summary, successful school improvement initiatives are school-based efforts that focus on a realistic number of priority target objectives that address critical

student needs and motivate school staff. These priority target objectives provide the focus for the building-level school-improvement activities, which can be integrated with the teacher-evaluation process.

It is important to note that other routes may be pursued to identify priority target objectives at the school building level. If a school system is not committed to district-wide educational planning, a planning process may be initiated at the building level to identify priority target objectives for school improvement. Also, at the secondary school level, including middle and junior high schools, there may be a need to identify priority target objectives for school improvement at the subject, team, or departmental levels, rather than at the school level. This approach is most appropriate in secondary schools.

The process that the principal, department head, or team leader uses to set a focus for school improvement is critical. We have all participated in planning processes where we felt like the pawns of the organization. The gears were to be set in motion and we were just one of the cogs. Setting the focus for school improvement in such an impersonal, mechanistic manner predestines the process to failure. In setting priority objectives for school improvement, the goal is to create a sense of purpose around which teachers are willing to coalesce, toward which staff want to work, and from which meaningful improvements in the quality of education will be provided to the students.

Principals involved in the leadership of school improvement teams would benefit from the perspectives provided by Peters (1987) in *Thriving on Chaos*. He makes it clear that the same principles of success prevail in the public as in the private sector. His management theory includes 45 prescriptions grouped into the following clusters (Peters, p. ix-x):

Creating Total Customer Awareness,
Pursuing Fast-Paced Innovation
Achieving Flexibility by Empowering People,
Learning to Love Change: A New View of Leadership at All Levels,
Building Systems for a World Turned Upside Down.

While initially it may take some thought to apply his prescriptions to schools, these prescriptions soon become the basis of a challenge for school leaders, the challenge to attain excellence in our schools. For example, "pursuing fast-paced innovation" challenges schools to continually experiment with new approaches to meeting student needs. Those approaches that are successful are continued, but in an atmosphere where still better approaches are sought. Within such an environment "flexibility must be achieved by empowering people." Teachers must have the liberty to identify critical school-improvement needs and to address these needs in small functionally organized groups.

Integrating School Improvement and Teacher Evaluation

Once priority target objectives for enhancing school effectiveness are identified, these objectives are integrated into the teacher-evaluation process by having teachers develop performance objectives that focus on the needs identified in these target objectives. In some school systems, this is done by simply asking teachers to consider the option of writing performance objectives that address critical school-improvement needs. A more effective approach would be to have teachers develop their objectives in school-improvement teams. First, the team would meet to select a school-improvement need of particular interest. Then teachers would develop individual performance objectives that address that need. Next, teachers would share and discuss their objectives with the team. Through such discussion, teachers would have the opportunity to refine and finalize their own performance objectives. Whether objectives are developed individually or in teams, it is important that teachers think through how school-improvement needs can be met and what their personal contributions might be.

Using the school-improvement-team approach to developing performance objectives has several advantages. First, it builds trust and collegiality. It creates an atmosphere where teachers can work together in developing and pursuing performance objectives, objectives which also serve as a basis for their evaluations. Second, it moves at least part of the teacher-evaluation process out of the principal's office and into the workplace. After teachers become accustomed to developing performance objectives in teams, this approach takes less time than if objectives were developed individually with the principal. Third, through this team approach, teacher evaluation has a more visible and significant impact on school improvement than if teachers worked alone on their performance objectives. Teacher evaluation becomes seen as something to improve schools, rather than to merely fill file cabinets. As Darling-Hammond (1986) notes:

A glimpse of what is possible is offered in the effective schools literature, which has discovered that participatory school management by teachers, based on collaborative planning, collegial problem solving, and constant intellectual sharing, can produce both student learning gains and increased teacher satisfaction and retention. Though teachers in these settings may or may not be involved in direct peer review for formal evaluation purposes, they nonetheless practice a form of evaluation and peer review when they identify problems, observe one another, share ideas, and ask, "How are we doing?" (p. 548).

The team approach to developing teacher-performance objectives can be illustrated by using a segment of the Windham Public Schools' improvement plan. This plan was developed through the Priority School District Program, a school-improvement initiative supported by the Connecticut State of Education (1987). Let us assume that the staff at one of the elementary schools identified a

priority objective dealing with the further development of the language arts curriculum guide. Also, these teachers decided that they wanted to address this objective by pursuing the following strategies listed in the improvement plan:

1. Need to include additional activities in the curriculum guide related to improving student writing.
2. Need to include additional activities in the curriculum guide which are compatible with the goals of the mastery test.

Teachers serving on the school-improvement team responsible for this objective would need to develop a set of complementary performance objectives that address each of these strategies. Three teachers on this team who are addressing the strategy related to improved student writing might develop a complementary set of objectives such as the following:

- Mr. Jones: To visit at least five school systems recommended by the Connecticut State Department of Education as having exemplary writing programs. The purpose of these visits is to identify additional activities for improving our students' writing.
- Ms. Smith: To participate in a three-day workshop on process writing conducted by the Connecticut Council of Teachers of English. The purpose of attending this workshop is to identify additional activities which might be used for improving our students' writing.
- Ms. Sage: To develop a system with Mr. Jones and Ms. Smith for (1) piloting the activities they have identified for improving our students' writing, and (2) revising the language arts curriculum guide to include those activities which are successful.

Integrating School Improvement and Teacher Evaluation with Staff Development

The quality of school improvement initiatives is enhanced when the teacher evaluation and school improvement processes are integrated with staff development. An effective staff development program is critical to the success of a school-improvement initiative as well as to the impact of a teacher evaluation process. As indicated in Figure 10.1, some school systems tend to organize the school improvement, staff development, and teacher-evaluation processes in a manner which is more disjointed than integrated. In school systems where the more disjointed approach is taken, teachers would develop objectives as members of school-improvement teams and would attend staff development meetings relevant to school-improvement needs. These same staff would develop perfor-

mance objectives for teacher-evaluation purposes and would become involved in staff development activities appropriate to those objectives. Given the limited time available for staff development in most schools, the quality of these separate inservice programs would tend to be mediocre at best. Thus, teachers' efforts to achieve those objectives developed separately for school-improvement and teacher-evaluation purposes would tend to have only a marginal impact on enhancing school effectiveness.

In school systems where school improvement, staff development, and teacher evaluation are integrated, teachers could develop performance objectives that both address school-improvement needs and serve as a basis for their evaluation. Also, these objectives could be addressed through the same staff development program. As indicated in Figure 10.1, the school improvement, staff development, and teacher evaluation processes complement each other when an integrated approach is taken and thus, could result in enhanced school effectiveness. As Murphy (1987, p. 160) notes, "One of the conclusions of the recent school improvement research is that schools work better when the parts fit together, when plans and activities are coordinated in a common effort to reach important school goals."

The need to deal with school improvement, staff development, and teacher evaluation an integrated manner is evident from the teacher comment below taken from a staff development survey.

Last week I was "inserviced" for my evaluator's objective. This week I'm going to a conference on my professional growth objective. Next week I'm going to be "staff developed" by the Priority Schools Project people. When are we going to start putting these things together so we can spend our time getting something meaningful accomplished for the kids in our building, rather than running from meeting to meeting?

Comments such as this one are not unusual. The time spent in many school systems on teacher-evaluation, staff-development, and school-improvement initiatives has increased substantially over the past five years, but the length of the school day and year have not changed significantly. These processes must be integrated. Otherwise, there is not sufficient time to pursue these processes as we work toward our goal of enhancing school effectiveness.

Putting the Purpose of School Improvement into Appropriate Perspective

The intent of this chapter has been to show that school improvement is an important purpose of teacher evaluation. The case is not being made for school improvement as the major or sole purpose of teacher evaluation. An effective teacher-evaluation process must address multiple purposes and be flexible

enough to address the purpose(s) most appropriate for enhancing the student learning experience in schools. At times this may mean a strong emphasis on school improvement. At other times the emphasis may be on enhancing professional growth. Sometimes a teacher may be in crisis and the focus needs to be on strengthening that staff member's performance. The issue is not whether the teacher-evaluation processes being used by a school is an accountability, professional-growth, or school-improvement model. None of these models is effective by itself. A teacher-evaluation process is effective to the extent that it is flexible enough to accommodate each of these purposes.

It is critical that principals exercise integrity when aligning the purposes of evaluation with the performance of the teacher and the achievement of that teacher's students. If a teacher has been given a reasonable period of time to strengthen his/her performance in the classroom, but is still performing unsatisfactorily, appropriate action must be taken. The focus of the teacher-evaluation process must turn from professional growth to intensive assistance and documentation for possible dismissal. While such decisions may not be easy to make, they are necessary as principals fulfill one of their basic ethical responsibilities — "Makes the well-being of students the fundamental value in all decision making and actions" (NASSP, 1973, p. 1).

While many educators agree that the teacher-evaluation process should have the flexibility to accommodate multiple purposes, including school improvement, they question whether they have the resources to implement such an approach. In school systems committed to effective teacher evaluation, sufficient resources are usually available to implement a process that has the flexibility to accommodate multiple purposes. The Teacher-Evaluation Cycle is such a process.

The Teacher-Evaluation Cycle

An overview of the Teacher-Evaluation Cycle can be presented in Table 10.1. It is important to note that this three-phase cycle is for tenured staff. New teachers and those not yet tenured are evaluated on a yearly basis using procedures similar to those described in the Appraisal phase.

Some Key Features of the Teacher-Evaluation Cycle

There are some key features of the Teacher-Evaluation Cycle, which are important to emphasize. First, the cycle begins with a strong accountability orientation by focusing on the intensive appraisal of teacher performance through the classroom-observation process. Recommendations are made at the conclusion of each formal classroom observation to strengthen or enhance the teacher's performance. Then teacher growth with respect to these recommenda-

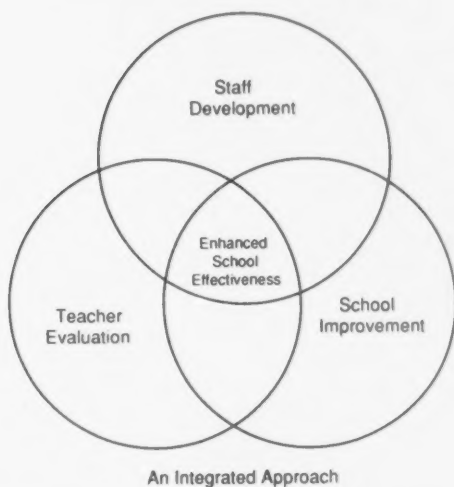
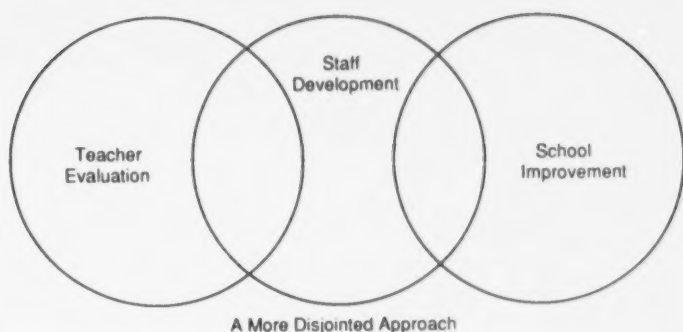


Figure 10.1 Approaches to Organizing the Teacher Evaluation, Staff Development, and School Improvement Processes.

tions is monitored through subsequent classroom observations and evaluation conferences. Classroom-observation conferences are opportune occasions for stimulating discussion about areas the teacher might address later in the evaluation process through performance objectives that focus on professional-growth or school-improvement initiatives. A teacher proceeds to the support phase of the cycle only if a favorable evaluation is received at the conclusion of the appraisal phase. This means the teacher has met the standards of performance expected by that school system. Each teacher is appraised at least every three years. If there is reason to believe that a teacher at either the Support or Continued Professional Development phases of the cycle is not performing satisfactorily, that teacher

may be placed in the Appraisal phase. In such situations the teacher must be informed of this placement in writing.

A second key feature of the Teacher-Evaluation Cycle is that it allows good teachers to assume more responsibility for their professional growth and evaluation. As the teacher moves into the Support or Continued Professional Development phases of the cycle, the evaluator provides support when needed as teachers pursue their professional-growth and/or school-improvement oriented objectives. In the Continued Professional Development phase, the teacher is no longer being evaluated by an administrator or supervisor. He or she is the evaluator. Sufficient time is available in the second and third phases of the cycle for teachers to accomplish meaningful outcomes and to reflect on where they are going professionally. Experience with school systems has indicated that teachers cannot always accomplish meaningful objectives during a year. Although substantial process toward objectives can be made over a year, it may take another year to refine and implement the products of such efforts. Since a reflective practitioner is a more effective practitioner (Schon, 1983), time for reflection is critical during the final phase of the cycle. Activities can be planned to help teachers to reflect upon what they have accomplished and to plan for the directions they would like to see their careers take in the future. As McLaughlin and Pfeifer (1988) note, "Reflection is a necessary first step in professional growth and improvement. . . . To the extent that a teacher evaluation stimulates a teacher to think about their practice, it can be a powerful force for self-improvement" (pp. 68-69).

A third key feature of the Teacher-Evaluation Cycle is that it documents a teacher's performance and development over a three-year period. Evaluation reports are completed during each spring of the cycle as follows:

- Year 1 — Appraisal Report,
- Year 2 — Progress Report on Objectives,
- Year 3 — Final Evaluation Report.

These evaluation reports are compiled into a portfolio which is included in the teacher's personnel file. Such a portfolio provides a richer perspective on a teacher's performance and accomplishments with respect to professional growth and/or school improvement objectives than the more traditional yearly evaluation reports.

A fourth and very important key feature of the Teacher-Evaluation Cycle is that it is manageable. Once the cycle is implemented, the evaluator works intensely on appraisal with nontenured teachers and only about one-third of the tenured teachers. Usually, this amounts to working intensely on appraisal with no more than eight to ten teachers, while either supporting or serving as a resource to other teachers who are in the second and third years of the cycle. Most administrators, supervisors, and department heads find that they have sufficient time to evaluate teachers effectively under this arrangement.

Table 10.1 An Overview of the Teacher Evaluation Cycle**1. Appraisal**

- Focus:** To conduct a thorough appraisal of teacher performance in light of the indicators of effective teaching.
- Events:**
- a. Fall conference to discuss and initiate the appraisal process.
 - b. Minimum of three classroom observations with follow-up conferences.
 - c. Spring evaluation conference to discuss the teacher's Appraisal Report and to develop objectives which can focus on:
 1. Strengthening performance with respect to the indicators of effective teaching.
 2. Professional growth initiatives.
 3. School improvement targets.

Orientation: A collegial, but accountability-oriented process where the evaluator assumes leadership for the teacher evaluation process.

2. Support

- Focus:** To support the teacher as work begins on objectives and to monitor progress in this regard.
- Events:**
- a. Fall conference to decide how the teacher's progress will be supported and monitored.
 - b. At least two evaluation conferences to support and monitor the teacher's progress.
 - c. Formal classroom observations and informal class visits as necessary.
 - d. Spring conference to discuss the teacher's Progress Report on Objectives.

Orientation: A collegial partnership, where the evaluator supports and guides the teacher's efforts to achieve his/her objectives.

3. Continued Professional Growth

- Focus:** To provide the teacher the opportunity to (a) pursue what needs to be done to attain his/her objectives and (b) reflect on where he/she is going professionally.
- Events:**
- a. Fall conference to review how the teacher's objectives will be attained and to discuss what strategies the teacher might use to reflect upon his/her performance.
 - b. Interim evaluation conferences as necessary.
 - c. Formal classroom observations and informal class visits as necessary.
 - d. In spring the teacher completes the Final Evaluation Report which is forwarded to the evaluator. This report includes a self-assessment of:
 1. the extent to which the teacher's objectives have been attained and
 2. those indicators of effective teaching which will provide the focus for his/her future professional development.

Orientation: A reflective process where the teacher assumes a more direct role in evaluating his/her performance and in setting a direction for future professional development.

*A Concluding Perspective on
Teacher Evaluation As School Improvement*

The emphasis on this chapter has been to affirm that school improvement is a critical purpose of teacher evaluation. When integrated effectively with staff development, teacher evaluation and school improvement lead to enhanced school effectiveness. The Teacher-Evaluation Cycle was presented as an example of an approach used by some school systems to address the multiple purposes of teacher evaluation, including school improvement.

Sergiovanni (1985) has discussed the relationship between the *mindscapes* of educational thought and the *landscapes* of educational practices. He comments that many educational improvement efforts are not successful, because the frames of reference that comprise the mindscapes of educational thought often are not appropriate guides for educational practice. This perspective is most appropriate for this chapter. A mindscape has been presented for teacher evaluation as school improvement. This mindscape is not presented as a prescription for teacher evaluation. It is presented as a beginning, as a starting point for reflection. Practitioners need to reflect upon this mindscape in light of the broad range of factors that have impact on the teacher-evaluation process and the quality of the educational programs in their school systems. Through such reflection a landscape will emerge for the teacher-evaluation process. Parts of this landscape will resemble the terrain covered in this chapter. Other parts may be very different. The similarity of this landscape to the mindscape of this chapter is not as important as whether that landscape enhances the student learning experience in schools. That is the goal of effective teacher evaluation.

References

- Bolton, D. L. (1973). *Selection and evaluation of teachers*. Berkeley, CA: McCutchen.
- Connecticut State Department of Education (1987). *Guidelines for the priority school district program*. Hartford, CT: Connecticut State Department of Education.
- Darling-Hammond, L. (1986). Teacher evaluation: A proposal. *The Elementary School Journal*, 86, 531-551.
- Iwanicki, E. F. (1981). Contract Plans: A professional growth-oriented approach to evaluating teacher performance. In J. Millman (Ed.), *Handbook of Teacher Evaluation*. Beverly Hills, CA: Sage Publications.
- Kappel, F. R. (1960). *Vitality in a business enterprise*. New York: McGraw-Hill.
- Kaufman, R. (1988). *Planning educational systems*. Lancaster, PA: Technomic Publishing Co., Inc.
- Loucks-Horsley, S. & Hergert, L. F. (1985). *An action guide to school improvement*. Alexandria, VA: Association for Supervision and Curriculum Development.
- McCune, S. D. (1986). *Guide to strategic planning for educators*. Alexandria, VA: Association for Supervision and Curriculum Development.
- McGreal, T. L. (1983). *Successful teacher evaluation*. Alexandria, VA: Association for Supervision and Curriculum Development.
- McLaughlin, M. W. & Pfeifer, R. S. (1988). *Teacher evaluation: Improvement, accountability, and effective learning*. New York: Teachers College Press.
- Murphy, J. (1987). Teacher evaluation: A conceptual framework for supervisors. *Journal of Personnel Evaluation in Education*, 1(2), 157-180.

- National Association for Secondary School Principal (1973). *A code of ethics for administrators*. Reston, VA: NASSP.
- Peters, T. (1987). *Thriving on chaos*. New York: Alfred A. Knopf.
- Purkey, S. C. & Smith, M. S. (1983). Effective schools: A review. *The Elementary School Journal*, 83(4), 427-452.
- Redfern, G. (1980). *Evaluating teachers and administrators: A performance objective approach*. Boulder, CO: Westview Press.
- Schon, D. A. (1983). *The reflective practitioner*. New York: Basic Books.
- Sergiovanni, T. J. (1985). Landscapes, mindscapes, and reflective practice in supervision. *Journal of Curriculum and Supervision*, 1(1), 5-17.
- Stiggins, R. J. & Bridgeford, N. J. (1985). Performance assessment for teacher development. *Educational evaluation and policy analysis*, 7(1), 85-97.
- Wise, A. E., Darling-Hammond, McLaughlin, M. W., & Bernstein, H. T. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: The RAND Corporation (R-3139-NIE).



PART II

Methods of Evaluation

11. Classroom Observation

Susan S. Stodolsky

Teacher evaluation, using almost any method, is complex and difficult. There are a number of possible approaches, but each has strengths and limitations. One common path is to obtain evidence by direct observation of teaching. In this chapter the author focuses on techniques for assessing teachers with direct observation. Specifically, the author considers what observations can and cannot contribute to the process of evaluating teachers, while discussing various issues associated with their use.

At the outset it must be clear that any method of evaluation is accompanied by a conception of what teaching is and, specifically, a conception of what good or effective teaching is. Yet it is not easy to find a consensus based on empirical evidence, theory, or values about the characteristics of good teaching or good teachers. In contemplating the use of an evaluation process, such as systematic observation, one is immediately confronted not only with methodological and procedural problems, but by the root problem of defining teaching itself.

Put another way, even if the user of an observational approach to teacher evaluation does not explicitly define good or effective teaching, the chosen method will instantiate some view of what is important in teaching. Therefore, the user of any observation system implicitly or explicitly accepts certain assumptions about the characteristics of teachers and teaching considered worth evaluating or judging.

Procedures are not neutral. They open up or constrict one's view of teaching and given teaching occasions. The categories built into an observation system focus attention on some aspects of instructional activity and exclude others. A view of teaching or instruction inheres in every system that is created. It is better to explicitly recognize these choices than to think that one is being objective, scientific, or neutral by using a systematic observational procedure. In developing or selecting an observational protocol, one embodies assumptions and beliefs

AUTHOR'S NOTE: Valuable research assistance was provided by Yu Amin Cheng during the preparation of this chapter. Tom Good and Jane Stallings provided suggestions on chapter content. Partial support was provided by the Benton Center on Curriculum and Instruction, University of Chicago.

about how teaching is constituted and makes decisions about those aspects worthy of attention and scrutiny.

The relation between one's belief about how teaching is constituted and the substance and form of observation can be illustrated by examining some actual observation procedures. Contrasting assumptions about teaching and accompanying differences in observation and evaluation procedures can be seen in specific cases.

As Wise, Darling-Hammond, McLaughlin, and Bernstein (1984) describe it, the Salt Lake City, Utah, district has established a teacher-evaluation system that incorporates a view of teaching as a professional activity. Thus, classroom instruction relies on teacher knowledge, experience, choice, and personal style. The district adheres to certain *criteria* of teaching but expects the specific ways in which the criteria are addressed in classrooms to vary across teachers. A variety of instructional approaches are assumed to be compatible with effective teaching.

For instance, one of Salt Lake City's criteria is "demonstrates appropriate strategies for teaching." This criterion is further defined by subcriteria such as, "adjusts techniques to different learning styles," "uses variety of techniques to teach specific skill or concept," and "demonstrates techniques that are appropriate to different levels of learning" (Wise et al., 1984, p. 14). This criterion and observations conducted to assess it rely on professional judgment and experience rather than a detailed specification of behavioral events to be observed. Depending on who and what is being taught, lessons involving discussions, lectures, group work, reading, and other instructional formats could all meet the criterion. In practice, it is school principals' responsibility to visit teachers' classes and evaluate them in light of district criteria, but they are not required to use a particular observation instrument.¹

The Salt Lake City program does not envision a classroom routine in which a prescribed set of steps are executed by every teacher nor does it require that specific behaviors be noted during an observation, but there are other evaluation protocols that come much closer to doing so. For instance, the Florida Performance Measurement System (FPMS) and other instruments aimed at assessing "generic competencies" contain specific behaviors thought indicative of effective teaching. In many instances, these instruments were developed from process-product research findings and took, as their definition of effective teaching, the characteristics of teaching found to correlate with student achievement. Less evaluator inference and judgment are required in the projection of those instruments than in others, such as those used in Salt Lake City or Greenwich, Connecticut (Wise et al., 1984).

Some items on the FPMS illustrate the more prescribed nature of teaching performance instantiated in this evaluation instrument. For instance, observers count occasions on which the teacher "conducts beginning/ending review, gives specific academic praise, provides for practice, states and applies academic rule, expresses enthusiasm verbally/challenges students, and stops misconduct"

(FPMS Summative Observation Instrument, n.d.). In general, the behaviors examined in the FPMS are weighted toward instruction that is teacher-directed and verbal, and away from student-centered instructional approaches. Therefore, the use of the FPMS carries with it a particular view of effective teaching that is not fully compatible with certain philosophical positions about learning and teaching. The FPMS may also be more suited to assessing instruction in certain areas of the curriculum, such as the skill topics in which the original process-product studies were constructed (Stodolsky, 1988).²

The author concludes this introductory discussion of the links between definitions of teaching and observation methods by noting two other related constraints. Different conceptions of teaching are more or less amenable to direct observation, and certain types of observational procedures are more or less compatible with different views of teaching. For example, the use of checklists of specific behaviors is compatible with the idea that teaching consists of certain classes of discrete teaching acts, but is not compatible with a view of teaching as preparing an environment for student learning.

Purposes of Evaluation: The Fit with Method

Because the purposes teacher evaluation can serve are discussed extensively in the previous section of this volume, it is not necessary to present them in detail here. However, the very close connection between purposes, goals, and methods must be explored in relation to observational procedures and what is known about their validity, reliability, and defensibility from a legal perspective.

A very general distinction can be drawn between evaluations that result in some definite decision regarding employment, certification, tenure, salary, or advancement, and those used for teacher development, assistance in learning to teach, or school improvement. The need to defend methods used in making personnel decisions about individuals is ordinarily more pressing, and the criteria more stringent than requirements for demonstrating the adequacy of methods used for improving teaching. While the appropriateness of such different standards is arguable, they must be considered when choosing among methods for evaluating teachers.

Differences in requirements for adequacy of methods are particularly germane with respect to the use of observation because it is very difficult to meet standards of validity, reliability, and generalizability with most currently used observation plans. Thus, for personnel decisions, some may find the use of observations alone indefensible. At the same time, it is difficult to imagine a plan for personnel decision making that did not include observation. Nevertheless, observations may be particularly well suited for the purposes loosely labeled "improvement," because they provide a base for discussion of actual classroom teaching by individual practitioners.

What Is Observation Good For?

Before discussing the specifics of how to conduct observations and what is known about their generalizability, validity, and reliability, one should consider what observation is especially good for and where it is less suitable. This matter relates directly to defining teaching and what can be captured by using direct observation.

In the most general terms, observation is good for looking at behavior or actions, but is a very limited tool for gaining understanding of thinking or feelings. Thus, if one wants to know how a teacher behaves with a group of youngsters during a lesson, observation may be a very good way to gather information. On the other hand, if one wants to know why individuals behaved the way they did or how (or if) they arrived at a plan to do so, observation alone would be a poor choice. Observation would also be an inadequate approach if one wanted to know how a person felt about the lesson while it was in progress or when it was complete.

Using observation for assessing teachers requires endorsing the idea that the target for evaluation can be *seen* and that it can be seen in a classroom lesson. It also assumes that visible actions provide an adequate appraisal of teaching competencies. A rather direct link between observable performance and underlying competence is assumed. The exclusive use of observation as an evaluation procedure presumes that observable, overt teaching actions provide a sufficient basis for judging a teacher's adequacy even though teaching is not just a set of observable performances or behaviors.

Much of what a competent teacher does appears out of public view. If observation is used as the only approach to teacher evaluation, certain aspects of teaching will not be reviewed. When using observation, one must assume that the behavioral sample adequately indicates competence in the unseen domains, that evidence regarding them is not central for evaluating teachers, or that evidence regarding unobserved competencies will be gathered in other ways.

Classroom observation leaves out direct systematic evidence about teacher planning, teacher assessment and modification of instructional materials, teacher choice and adaptation of instructional methods, and teachers' working relations with colleagues, parents, and members of the school community. In some cases, the information value of observations depends on the type of instructional events observed and the ability of observers to evaluate the quality of subject matter content actually taught. In particular, direct observation of classroom instruction may not be consistently informative about such domains as teachers' pedagogical knowledge of subject matter (Shulman, 1986), teachers' content mastery of the subjects they teach, and teachers' assessment and feedback to students.

Just as observation is particularly suited for assaying behavior and actions in general, specific types of observation systems are suited for recording and analyzing certain classes of teaching behaviors and actions. Evaluators must decide whether to use observation as an exclusive method of teacher evaluation

or to have observations play some role in light of their purposes and goals. Similarly, specific types of observations must be weighted for their match with an evaluator's purpose and beliefs.

In the remainder of this chapter, some types of observation systems are examined and what is known about their validity and reliability is considered. The author also examines what is known about stability of performance when teaching is assessed with observation instruments. Particular attention is paid to the role of context in teaching and the issues of cross-subject and cross-occasion generalizability. After these general considerations, some procedural matters are discussed. Included are how to plan and select occasions for observations, who observers should be and how they should be trained, issues of due process and notification, and pre- and post-observation conferences. The chapter concludes with some recommendations regarding the use of observation for different kinds of teacher evaluations.

Observation Systems

Observation as a method for evaluating teaching can take many forms. Until fairly recently, teacher evaluations were often made by a school principal or supervisor, based in part on one or more relatively short observations during instruction in a teacher's classroom. The procedures for observation were typically "ad lib" or informal — no special format was required. The observer reflected on what was seen and either made a summary judgment on rating of the teacher or made suggestions about instructional improvement.

More recently procedures have been instituted in many school districts, municipalities, and states that involve systematic observation of teachers using specific protocols and instruments (Kowalski, 1978; Darling-Hammond, Wise, & Pease, 1983; Wise et al., 1984). A broad range of instruments including behavior checklists, category systems, narrative records, summaries, and rating systems is in use. Procedures vary also as to whether they require high or low levels of inference to categorize observed behaviors.

Good references about instruments and observational procedures are available to the interested reader. Sources for compilations of instruments and discussions of the characteristics of different procedures include Evertson and Holley (1981), Evertson and Green (1986), Simon and Boyer (1967), Kowalski (1978), Borich and Madden (1977), Medley, Coker, and Soar (1984), and Stodolsky (1984).

In this section a brief overview of observation systems is presented, with particular attention to the contribution each can make to the teacher evaluation process. A general distinction between open and closed observation methods is useful (Wright, 1960). In open systems, in which the author includes ad lib notes, narratives, specimen records as developed by Barber and Wright (1955) and Wright (1967), as well as films or videotapes, the observation record contains a

description of a sequence of behavior in ordinary language or mechanically reproducible form. In taking open records, an effort is made to fully describe "all" behavior as it occurs and to avoid interpretation or selection.

Open systems are sometimes used in connection with summary ratings or judgments. A narrative or specimen record provides a rich description of classroom events that can be used to make evaluations or assessments. Open records also provide specific examples that can be discussed or analyzed in a conference between an evaluator and teacher or in a peer review.

In contrast, closed systems are collective and focus on specific types or aspects of behavior. Closed systems include category and sign systems, as well as behavior checklists and rating scales. The typical category system contains mutually exclusive and exhaustive categories that are applied to behavioral events of interest. When using a category system, every relevant observed behavior is placed in one and only one category. Good and Brophy (1987) provide examples of category systems, such as their Form 9.2 for assessing student involvement in assigned work. In its application, they suggest that every student be observed and their behavior placed in one of four mutually exclusive categories every two minutes. The categories are: "clearly involved in assigned work; can't tell — may be thinking; definitely not doing assigned or chosen work; and misbehaving" (p. 400). The Flanders Interaction Analysis System and the Observation Schedule and Record (OScAR) are other familiar examples of category systems.

Sign systems also contain predefined coding categories, but while the categories are mutually exclusive, they are not exhaustive. Only specific selected behaviors are recorded when using a sign system; the observer ignores behaviors that do not fit in any category and are therefore not of interest. The Classroom Observations Keyed for Effectiveness Research (COKER) is an example of a sign system. Another example from Good and Brophy (1987, p. 520) relates to the type of question asked by a teacher. Their form provides three categories of questions: academic, factual; academic, opinion; and nonacademic. However, some questions cannot be accommodated with these three categories, for example, those which require application, synthesis, or other higher mental processes and are academic but not factual. The selection of concern categories for attention, as in this case, is characteristic of a sign system.

Behavior checklists are a common form of closed observation used in teacher evaluations. Checklists are set up so that the observer indicates the presence or absence of a given behavior during a lesson or during a specified time interval in a lesson. Frequently, short time intervals such as 30 seconds to one minute are established and the occurrence of the coded behaviors is indicated for each interval. A number of state-mandated observation systems, such as the Florida Performance Measurement System and an instrument developed in North Carolina (White, Wyne, Stuck, & Coop, 1987) are behavior checklists.

Rating systems, in which scaled judgments of performance are made, are discussed by Good and Mulryan in this volume. In contrast to many closed

systems, especially checklists, ratings represent a high-inference procedure in which the evaluator makes summary judgments about a teacher's performance. Examples of ratings would be judgments regarding the degree to which a teacher is an effective manager or is supportive of student emotional development. As mentioned earlier, ratings are sometimes produced after systematically recording open observations as well as after use of behavior checklists.

The author has already made passing reference to high-inference and low-inference observation systems. The distinction concerns the specificity (and usually narrowness) of behavioral definitions in an observation system. In low-inference systems, such as the Florida Performance Measurement System, specific and easily identified behaviors are codified. Observations are usually directed at the frequency of discrete teaching acts. Consequently, observers usually reach high levels of accuracy with low-inference systems unless they are asked to record excessive numbers of behaviors in short intervals. A strength of low-inference systems is their reliability. High-inference systems require more judgment on the part of the observer and usually relate to bigger patterns of behavior and events than low-inference systems. Consequently, high-inference systems usually demand more evaluator training than low-inference systems, and may suffer from lower reliability. (But c.f. Shavelson and Dempsey-Atwood, 1976, who found high-inference ratings more generalizable and stable.)

A trade-off between validity and reliability is almost always present in selecting either a low- or high-inference system. If one believes that teaching is decidedly more than the sum of discretely definable parts, a low-inference system, no matter how reliable, is likely to be unacceptable in terms of its validity. Instead, observations must be made of whole activity structures created in classrooms and more must be known about the meaning of behaviors. A high-inference system is a likely result. The state of Oklahoma (Oklahoma State Department of Education, 1986), uses a high-inference system in its assessment of residency teachers, by employing observation guidelines that specify general domains such as human relations, teaching, and pupil assessment. Oklahoma provides teams to assess new teachers and relies on team members to use their professional expertise in following the guidelines. The Salt Lake City procedure described by Wise et al., (1984) and discussed earlier is another example of high-inference observation.

In this writer's view, there has been an unfortunate tendency to excessive concern with reliability and objectivity in teacher observation systems accompanied by insufficient concern about validity and definitions of teaching. Instead, I think selection of a system should be determined by the match between the school district's view of teaching and the adequacy with which the instrument reflects that view.

Stability, Generalizability, Reliability and Validity in Teaching

The interrelated constructs of stability, generalizability, reliability, and validity have all been studied with respect to teaching behavior and teaching performance, and the conclusiveness of the research on these topics varies. The constructs themselves are closely related and only central issues will be highlighted here.

These properties of teaching and its assessment are central to a consideration of observational methods because they suggest the extent to which evaluative inferences made from observations are likely to be trustworthy, fair, and defensible. They also provide guidance about how many observations are necessary for evaluation purposes and the conditions under which observations should be made.

The literature examining issues of reliability, stability, and generalizability has addressed the way in which measurement conditions and situational and contextual factors influence these features of observations. Reliability of measurement as used here refers to the objectivity and reproducibility of an observation on a given occasion. If a procedure is reliable in this sense, the summary scores, frequency counts, or ratings derived from an observation occasion should be highly similar when different observers are involved. Measurement conditions known to influence reliability or interobserver agreement include observer training and experience, whether instruments require high or low inferences, and clarity of instrument categories. Lack of reliability is often a function of unclear definitions, the need for difficult observer judgments or inferences, insufficient training, observer fatigue, an overly complex or demanding observation protocol, or behavioral complexity. Most measurement conditions that influence reliability can be systematically addressed with corrective action.

The aspect of reliability more akin to consistency, stability, or generalizability represents the idea that similar scores, ratings, or judgments will be produced when the same teacher is observed on a number of occasions. More fundamental, these constructs rest on the assumption that teaching behavior itself is stable and consistent. In other words, stability, generalizability, and consistency require teaching to look quite alike across contexts and occasions. For this form of reliability to be achieved, teachers must be rather consistent in the way they teach over time and the way they teach different groups of students.

Available empirical literature has documented a variety of contextual and situational factors that limit the stability, generalizability, and consistency of teacher behaviors and teacher effects. While there are some exceptions, in general, consistency and stability are not found in the teaching domain. Regardless of the type of observations recorded, certain major contextual factors are associated with variation in teaching that produces instability, inconsistency, and a lack of generalizability.

Subject matter, grade level, lesson type, and lesson goal account for systematic variation in teaching behaviors and instructional arrangements (Stodolsky, 1988; Stodolsky, 1984; Shavelson & Dempsey-Atwood, 1976; Shavelson, Webb, & Burstein, 1986). Reasonably convincing evidence suggests that teachers vary their instructional approach depending on what they are teaching (e.g., math, reading, or social studies), the grade level of pupils, the instructional lesson type (e.g., recitation, lecture, seatwork, groupwork, reading circle, or lab experiment), and the lesson goal (for example, factual mastery, problem solving, or enhancement of self-esteem). In addition, some have suggested (Wineburg, in prep; Leinhardt, 1987) that the specific topic of instruction (e.g., American history or social problems) and certain characteristics of pupils (Barr & Dreeben, 1983) also affect instructional practice. A teacher's philosophy, district policies, and other individual and institutional factors are associated with variation in instruction, too (Schwille et al., 1983).

The type of observation and the judgments made from it also relate to degree of consistency or stability, although this topic needs much more study. It appears that very global judgments of competence are reasonably stable (Shavelson & Dempsey-Atwood, 1977). On the other hand, Calkins, Borich, Pascone, Kugle, & Parston (1978) found inconsistent results and evidence of instability when a variety of high- and low-inference systems were applied to three observations of the same teachers teaching the same subject matter in junior high school lessons. Few studies have found high levels of consistency or stability of performance on the part of teachers, whether observing discrete, molecular behaviors and taking frequency counts, assessing student involvement or achievement, or making judgments of managerial and other instructional skills.

Sampling Occasions

The typical evaluation process involves a small number of observations of a given teacher. By current practice, two or three one-hour observations represent the high end of time spent observing an individual (Kowalski, 1978; Wise et al., 1984; McLaughlin & Pfeifer, 1988). A traditional measurement perspective would suggest that increasing the number of observation occasions would increase consistency and reliability. Rowley (1978) demonstrated that more observations do produce higher reliability on some classroom measures, particularly when the circumstances under which the observations are taken are relatively similar. However, the relationship between the number of observations and reliability is not a simple linear one, and increasing occasions of observation increases reliability in different degrees, depending on the particular measure being assessed. Rowley (1978) demonstrated that more observations of shorter duration are more likely to produce better reliability than a few, longer observations even though the amount of total time observed is equal in the two cases. Intuitively, it may seem better to watch intact lessons rather than shorter parts of

lessons, but from the point of view of reliability, shorter observations probably produce better evidence.

Perhaps more pivotal in this discussion is that situational and lesson-type variation implies that simply increasing the number of observations of a teacher will not increase consistency or reliability. Because instruction itself is variable, more observations alone will not obviate the intrinsic variability. Thus, Rowley (1978) and Calkins et al., (1978) as well as Darling-Hammond, Wise and Pease (1983) and Stodolsky (1984) do not recommend increased numbers of observations per se as a solution to the instability problem.

More observations are much more likely to help produce stable and generalizable results if factors known to produce systematic changes in instruction are used to produce a sampling plan. For instance, a series of observations of particular lesson types in a particular subject (e.g., math recitations) might result in more reliable and internally consistent findings about an individual teacher than would a similar number of observations of the teacher taken across the curriculum without a deliberate sampling strategy. However, observations of math recitations would not necessarily generalize to the teacher's behavior during seatwork, lectures, or other activities. Nor would such observations necessarily be informative about the teacher's instructional abilities in other subject areas. To produce generalizability and validity as well as reliability, samples of central lesson types in a number of subjects would have to be observed.

The author concludes that teaching is context dependent and stability is likely only within well-defined contexts, such as lesson types within subjects and grade levels. Given this degree of specificity of teaching behavior, a number of practical alternatives for collecting observations must be considered. Reviewing the purpose for observation will facilitate selection among various options for sampling. In a summative personnel decision, an effort to obtain generalizable results for a teacher seems desirable. When observations are used for purposes of improving instruction, generalizability and stability may be less prominent concerns. A single observation or a small number of observations in connection with almost any lesson can produce profitable suggestions and discussion, as long as it is clear that a complete view of a teacher's abilities has not been obtained.

A sampling plan encompassing a number of observations selected to represent the key contexts in which a teacher works represents a potentially desirable, if expensive, approach. Another alternative is to focus on a small number of the most frequent lesson types used or subject matters taught by a teacher. A third strategy would select lesson types thought indicative of the teacher's abilities because of their importance, centrality, or relevance. Each of these sampling plans takes into account main sources of variation in teaching and accommodates them in different, but plausible, ways.

A candid acknowledgment that a small number of observations cannot be representative of a teacher's abilities is another solution to the sampling problem. Under this scenario, a teacher could be asked to prepare a showcase lesson,

tantamount to an "existence proof" with respect to his or her teaching competence. A specially prepared lesson does not tell observers what a teacher usually does, but it does demonstrate what a teacher can do. For certain evaluation purposes, such as observing during a period of training or remediation, a teacher's performance under showcase conditions may provide useful evidence.

Simulations represent another approach somewhat related to showcase lessons. Simulations are addressed to producing comparable conditions of measurement across teachers. In a simulation, teachers are asked to teach outside their own classes. The lesson focus and the pupils to be taught along with available materials are usually stipulated. A strength of the simulation procedure is that it provides a lot of control over the circumstances and demands placed on a teacher while observations are taking place. However, the validity and generalizability of measures obtained in a simulation with respect to a teacher's usual classroom circumstances may be poor.

The author has suggested a number of strategies for collecting observations to address the problem of lack of generalizability and stability in teaching. Perhaps belatedly, it is necessary to highlight the importance of validity in all these considerations. To have a valid observation procedure, both the instrument itself and the sampling plan with which it is used, must adequately reflect the definition of teaching held by the school district or other evaluation agent. Earlier, the author discussed the issue of defining teaching in connection with instrument development, but it is also the central issue with respect to validity.

As instruments and sampling plans are selected, they must be weighed for the adequacy with which they match an adopted conception of teaching. For example, behavior checklists incorporating several specific indicators of teaching competencies, such as those used in Georgia, Florida, and Tennessee, are based on a view of teaching that deemphasizes the role of contextual and personal factors in teaching. In applying these instruments, it is considered appropriate and valid to observe almost any lesson taught by a teacher. For those who do not define teaching as discrete acts, however, both different instruments and different sampling strategies would be required.

If practical limitations make it unfeasible to produce a comprehensive view of teaching through observation, other data collection methods should be employed to round out the picture of a teacher's competence. Written materials including lesson plans and student work, as well as conversations with teachers, can provide useful information. The fact that the role of observation may be limited in an evaluation process does not diminish its utility. In light of the flexible and adaptive nature of teaching, a fully satisfactory way to use observation (i.e., to meet validity, reliability, and generalizability requirements) may not always be available. Users must accept the limitations of observations as sources of evidence about teaching while recognizing that they provide a needed direct view of teaching processes in action.

Some Procedural Considerations

When a teacher is to be observed as part of an evaluation, certain procedural and practical matters arise. First is whether the teacher is informed in advance that he or she will be observed. Prior notification is sometimes a contractual obligation, and in many cases seems desirable. A planned observation permits a preobservation conference in which teacher and evaluator can discuss the intent of the lesson, as well as observation procedures and expectations. Planning observations also permits more precise selection of the types of lessons to be observed. On the other hand, unannounced observations may locate unprepared, abusive, or unmotivated teachers who alter their usual behavior when they know they are going to be observed.

Another issue is whether or not the teacher should be familiar with the observation instrument and the way in which judgments are derived prior to being observed. Will such knowledge produce a test-preparation mentality similar to that engendered in students preparing for national exams or is it more analogous to describing the general content of a test to an examinee? No research-based answer to this question is available, although there is anecdotal evidence that some teachers try to produce specific targeted behaviors when low-inference checklists are in use. In this writer's view, teachers should know the standards and expectations by which they will be evaluated. If the process of evaluation influences teachers to teach in certain ways, it is the evaluator's responsibility to consider whether such an influence is desirable. If an evaluator is troubled by this possible consequence, the observation procedure itself may be inadequate in its representation of the teaching process.

A conference following observation to discuss the observation and judgments made by the evaluator is almost always necessary. Teachers have the right to know when conclusions have been drawn about their teaching performance and to plan their goals and reasons for observed instructional activities. Such conferences may also help an evaluator make the most appropriate and informed recommendations regarding the teacher by providing contextual information and the teacher's rationale for the observed instruction. Of course, when observations take place as a means of sharing information and perceptions with colleagues, postobservation discussions are naturally part of the process.

Who Should Evaluate?

The assignment of the evaluation role to individuals within a school system or state is governed by tradition, job descriptions, and occasionally legislated. Within school systems, observers are usually superintendents, principals, assistant principals, department chairs, or curriculum supervisors. Faculty from colleges and universities often recommend that the state grant provisional credentials when a teacher completes training at their institution of higher education.

Many beginning-teacher evaluation programs involving residencies and internships use principals and curriculum supervisors as evaluators. State programs for assessment of teachers who already hold licenses are conducted by the various personnel already mentioned, as well as by cadres of teachers trained as evaluators. Increased interest in peer review and professionalization has given a more prominent place to teacher colleagues in the evaluation process.

The purpose of an evaluation is relevant in thinking about who will conduct observations and make judgments about teachers. Standards of objectivity, fairness, and equity are important under all circumstances and must be achieved through the exercise of professional conduct and appropriate training. However, when teachers are to receive specific feedback meant to improve their teaching, it may be desirable that the evaluator have expertise similar to that required in the teaching setting.

It is probably easiest for teachers who instruct similar grade levels and subjects, department heads, and curriculum supervisors to make instructional suggestions to a colleague whose assignment matches their own rather closely. A number of teacher-evaluation programs, for example the Oklahoma Teacher Reform Act of 1980, require the inclusion of individuals who are teaching in similar circumstances as members of an evaluation team. The general principle of peer review has been adopted in many professions outside teaching with salutary results (Wise & Darling-Hammond, 1988).

A close match in assignment and expertise may be especially important when an observation protocol involves more open methods and requires high rather than low inferences. Under such circumstances the evaluator must be able to make use of contextual and pedagogical information and draw on relevant experiences in doing so. It is likely that a high school English teacher would be a helpful observer in another English teacher's classroom but might have much less to offer a colleague who taught mathematics or science.

Because specific subject matter knowledge and pedagogical practices differ across teaching settings, on occasion it may be desirable to have more than one observer. Especially when a team is used, teachers might be given the opportunity to suggest a teaching colleague as a team member. Discussions that incorporate a number of perspectives about a given teaching occasion would probably be an optimal way to promote improvement of teaching and professional regard, as well as sustain and develop professional values and norms.

A highly structured checklist might more readily be put in the hands of a trained evaluator whose expertise was different from the observed teacher. When the observation protocol is tightly defined, prior knowledge of particular types of teaching may be less important, although an educational background still seems necessary. Nevertheless, similar background and assignment should enhance the validity of an observation even with a closed category system.

Observer Training

As has been true for a number of topics discussed in this chapter, suggestions regarding observer training are influenced by the specific type of observation and its purpose. In the case of learning to use a category system or checklist, observers must learn definitions and practice applying the instrument to a variety of teaching situations. For an instrument of any complexity, this kind of training typically represents an investment of a number of days and can extend over two to three weeks. Training is considered adequate when standards set for reliability and objectivity are met. As indicated earlier, standards vary depending on the purpose of the observation and, to some degree, the extent to which it will influence the livelihood or future employment of the teacher.

Training in the use of both closed and open methods can frequently be conducted with the use of video tapes rather than observations in actual classrooms. By training observers with videotapes, a judicious selection of teaching behaviors and situations can be made. In this way, observers can be exposed to a range of teaching situations similar to those they will encounter while actually conducting evaluations. In addition, however, it is important that observers actually go into classrooms and carry out "live" observations so that they can become familiar with the process of focusing their attention on relevant events in the context of a total classroom setting. Actual classroom observations also provide training in face-to-face dealings with teachers, students, and other school personnel.

Perhaps one of the more difficult aspects of evaluator training is the specification of means by which observation records, notes, or checklists are translated into usable summary judgments. In some cases, numerical standards relating to the frequency of certain behaviors are specified as "satisfactory" and the procedure is rather mechanical. However, many protocols, for example those requiring the formulation of recommendations for improvement of instruction, require the exercise of professional judgment and expertise.

Workshops or meetings in which evaluators discuss standards and expectations and acceptable views of teaching seem essential in using many instruments. Evaluators must make an effort to develop common approaches to the assessment process and its resultant judgments. Because different types of instruction are defensible, diversity may be built into the evaluation criteria. However, standards and expectations should be made explicit and available to teachers, to the extent that they can be articulated.

The Role of Observation

The omission of direct observation from evaluation of teachers seems unthinkable. How can one evaluate teachers without seeing them in action? At the same time, observations should not be given so prominent a role that many

important unobservable aspects of teaching are ignored. Perhaps an ideal teacher evaluation would incorporate multiple sources of evidence about teachers and the effects they have on students.

Pre- and post-observation interviews can provide some of the additional information needed to round out a picture of teaching competencies. Interviews can reveal teachers' plans, as well as their reasons for constructing instruction as they do with a particular group of students. In addition to interviews, certain paper and pencil instruments may be used to assess teacher knowledge of subject matter and pedagogy. Evidence about student learning and attitudes, as well as perceptions of parents and other members of the community may also be helpful.

Because instruction is an interactive and adaptive process, it would be too much to expect a clearcut process by which to observe and evaluate teachers. On the other hand, a systematic approach coupled with some modesty is feasible and essential. The state of the art in using observations will probably improve to some extent, but it is limited largely by our understanding of the constituents of teaching itself. Value choices about the ingredients of good teaching will still remain, even as progress is made in understanding teaching. Professional judgment and insight are likely to retain key roles in the evaluation process. Nevertheless, systematic and fair procedures can be enacted to facilitate professional judgments. Classroom observations are likely to be the centerpiece of a systematic evaluation strategy.

Notes

1. Discussing the system is not meant as an endorsement of it nor as documentation that it is actually implemented as described.

2. The validity of inferring the characteristics of effective teaching from process-product studies could be discussed at length and can be questioned on methodological and philosophical grounds. An extended treatment of some of the pitfalls of this approach can be found in Stodolsky (1988). In any case, it represents a particular strategy that clearly reveals the link between a conception of what effective teaching is and observational procedures.

References

- Barker, R.G., & Wright, H.F. (1955). *Midwest and its children: the psychological ecology of an American town*. New York: Harper & Row, Publishers.
- Barr, R., & Dreeben, R. (1983). *How schools work*. Chicago: University of Chicago Press.
- Borich, G. D., & Madden, S.K. (1977). *Evaluating classroom instruction: A sourcebook of instruments*. Reading, MA: Addison-Wesley.
- Calkins, D., Borich, G.D., Pascone, M., Kugle, C.L., & Parston, P.T. (1978). Generalizability of teacher behaviors across classroom observation systems. *Journal of Classroom Interaction*, 13(1), 9-22.

- Darling-Hammond, L. (1986). A proposal for evaluation in the teacher profession. *Elementary School Journal*, 86(4), 531-51.
- Darling-Hammond, L., Wise, A.E., & Pease, S.R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, 53(3), 285-328.
- Evertson, C.M., & Green, J.L. (1986). Observation as inquiry and method. In M.C. Wittrock (Ed.) *Handbook of research on teaching*, (3rd ed.) New York: Macmillan Publishing Company.
- Evertson, C.M., & Holley, F.M. (1981). Classroom observation. In J. Millman (Ed.), *Handbook of teacher evaluation*. Beverly Hills, CA: Sage Publications.
- Good, T.L., & Brophy, J.E. (1987). *Looking in classrooms*, (4th ed.) New York: Harper & Row, Publishers.
- Kowalski, J.P.S. (1978). *Evaluating teacher performance*. Arlington, VA: Educational Research Service.
- Leinhardt, G. (1987). *Integration of lesson and subject matter knowledge in teaching*. Paper presented at the annual meeting of AERA, Washington, CA.
- McLaughlin, M.W., & Pfeifer, R. S. (1988). *Teacher evaluation: Improvement, accountability, and effective learning*. New York: Teachers College Press.
- Medley, D.M., Coker, H., & Soar, R.S. (1984). *Measurement-based evaluation of teacher performance: An empirical approach*. New York: Longman.
- Oklahoma State Department of Education (1986). Oklahoma teacher reform act of 1980. (12 p.).
- Rowley, G. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. *Journal of Educational Measurement*, 15(3), 165-180.
- Schwille, J., Porter, A., Belli, G., Floden, R., Freeman, D., Knappen, L., Kuhs, T., & Schmidt, W. (1983). Teachers as policy brokers in the content of elementary school mathematics. In L.S. Schulman & G. Sykes (Eds.), *Handbook of teaching and policy*. New York: Longman.
- Shavelson, R.J., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, 46(4), 553-611.
- Shavelson, R.J., Webb, N.M., & Burstein, L. (1986). Measurement of teaching. In M.C. Wittrock (Ed.) *Handbook of research on teaching*, (3rd ed.) New York: Macmillan Publishing Company.
- Shulman, L.S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4-14.
- Simon, A., & Boyer, E.C. (1967). *Mirrors for behavior: An anthology of classroom observation instruments* (6 vols). Philadelphia: Research for Better Schools.
- Stodolsky, S.S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher*, 13, 11-22.
- Stodolsky, S.S. (1988). *The subject matters: Classroom activity in math and social studies*. Chicago: University of Chicago Press.
- White, K., Wyne, M.D., Stuck, G.B., & Coop, R.H. (1987). Assessing teacher performance using an observation instrument based on research findings. *NASSP Bulletin*, 71(497), 89-95.
- Wineburg, S.S. (in prep.) *Historical problem solving: A study of the cognitive processes used in the evaluation of documentary evidence*. Ph.D. dissertation, Stanford University.
- Wise, A.E., Darling-Hammond, L., McLaughlin, M.W., & Bernstein, H.T. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: The RAND Corporation.
- Wise, A.E., & Darling-Hammond, L. (1988). *Licensing teachers: Design for the teaching profession*. Santa Monica, CA: The RAND Corporation.
- Wright, H.F. (1960). Observational child study. In P.H. Mussen (Ed.) *Handbook of research methods in child development*. New York: Wiley.
- Wright, H.F. (1967). *Recording and analyzing child behavior*. New York: Harper & Row, Publishers.

12. Teacher Ratings: A Call for Teacher Control and Self-Evaluation

Thomas L. Good

Catherine Mulryan

Introduction

This chapter begins with a brief history of the use of rating forms for teacher evaluation, illustrating that rating forms are associated with enduring problems that have no simple solutions. In particular, external evaluation has not produced agreed upon procedures for evaluating teachers. In the second section the authors highlight those historical and contemporary influences that serve to make teacher evaluation a problematic activity.

In the third section of the chapter, the authors support the use of descriptive scales to provide teachers with information that encourages them to improve their teaching,¹ arguing that such information is valuable if it encourages self-reflection, analysis, and growth. Rating instruments — and how they are used — should be largely controlled by teachers. This is especially the case after teachers have demonstrated general competency in managing the classroom. The classroom context and local goals are extremely important considerations in the development of appropriate instruments.

Teacher Evaluation: A Historical Perspective

In this section, the authors examine a range of issues related to the development and use of rating scales for the evaluation of teachers. Rating scales were first used in the late nineteenth century. By the 1950s, several thousand studies

AUTHOR'S NOTE: The authors acknowledge the support provided by the Center for Research in Social Behavior at the University of Missouri-Columbia. We also wish to thank a number of individuals who have provided helpful advice to us during manuscript preparation. These individuals include Jere Brophy, Howard Ebmeier, Gail Hinkel, Teresa Hjelming, and Richard Manatt.

using rating scales had become available in the teacher effectiveness area (Dunkin & Biddle, 1974, p. 12). During this period the monitoring and evaluation of teachers were important features of educational administration and practice (Barr, 1931; Morsh & Wilder, 1954). The school survey movement, standardized testing, and the increasing bureaucratization of public school administration that characterized the early decades of the twentieth century were largely responsible for the study of teaching effectiveness, teacher appraisal, and concern with teaching standards. An additional stimulus for teacher appraisal was the increasing demand for teacher accountability. Marked increases in teachers' salaries in the early decades of the twentieth century and growth in the status of the teaching professional created public interest in obtaining value for funds spent.

Early Research on Teacher Effectiveness

The period roughly from the beginning of this century to the 1930s was characterized by increasing interest in teacher effectiveness, by the development of numerous approaches to the appraisal of teaching, and by the creation of a variety of evaluation instruments. The earliest research on teacher effectiveness used students' observations to identify effective teacher traits. Several such studies appeared between 1896 and 1955; Kratz's (1896) study, "Characteristics of the Best Teachers as Recognized by Children," was the pioneering work in this area. Interest in student observation and opinion waned in the early part of the twentieth century (Beecher, 1949; Morsh & Wilder, 1954) only to be revived to some degree in the late 1930s.

Studies that identified traits of effective teachers as determined by experts, including administrators and professors of education, began in 1915 with a study by Anderson (1917). The Commonwealth Teacher Training Study (Charters & Waples, 1929) pushed this type of research as far as it could be taken (Medley, 1972). An approach to the study of teacher effectiveness that is similar in many ways to teacher trait studies is what Beecher (1949) calls *the negative approach*. Such studies sought to determine the reason why teachers failed and produced lists of perceived teacher weaknesses (e.g., Littler, 1914).

Rating scales became popular in research and teacher evaluation in the early decades of the twentieth century. The first rating scale appeared around 1915 (Elliott, 1915), and by 1930 several hundred such devices were available (Medley, 1972; Morsh & Wilder, 1954). Whereas the earliest ratings of teachers were done mainly by administrators, pupil ratings and teacher self-ratings were in the use by the late 1930s. However, by the mid-1950s through contemporary practice, the emphasis has returned to administrative evaluation of teachers.

The formation of the Committee on Child Development of the National Research Council in 1920 resulted in the creation of the first observational development research instruments, which was an important step in teacher-effectiveness research. The pursuit of greater objectivity in these instruments, however, frequently resulted in the collection of relatively meaningless data, and

significant improvements in observational instruments did not occur for at least another 25 years (Medley, 1972).

The emergence of the Ohio Teaching Record (1941) was a major step forward in the study of teacher effectiveness. This instrument encouraged a cooperative approach to the supervision of teachers in which teacher and supervisor worked together to improve teaching. It was a diagnostic instrument that focused on how well students' needs were being met by teachers in the classrooms. The cooperative approach to teacher supervision was further stimulated by a report by Troyer and Pace (1945) for the American Council of Education. This report deals with teacher evaluation in general and also addresses the use of techniques for cooperatively evaluating teaching, including the Ohio Teacher Record. Baxter (1941) focused on pupil-teacher interaction as a basis for estimating teacher effectiveness, and in so doing highlighted a factor that had previously been largely neglected.

Rating Scales

In the literature the terms rating scale, score card, checklist, and appointment blanks all refer to similar rating instruments that were the most popular form of teacher appraisal up to the late 1950s and early 1960s. A wide range of rating instruments were produced, "from the 'one-man' products of the 'all-wise' superintendent to those resulting from careful research" (Beecher, 1949, p. 32).

A fairly typical rating scale, or score card, was that of Elliot (1915), which was presented at a conference of city school superintendents in Washington D.C. This scale included 42 traits that had been selected from previous research and were considered essential for effective teaching. Judgments were presented in numerical form and a total rating was obtained by summing the scores for each trait. A second type of rating scale produced a qualitative rather than a quantitative description of teaching and generally utilized a 3-to-5-point scale. These scales, called *quality scales* by Barr (1931), required rater judgment with regard to the presence or absence of traits or the degree to which they were present. According to Beecher (1949), this approach is typified in a score card produced in 1915 by the New York Board of Municipal Research.

A third type of rating scale (e.g., Rugg, 1920) was intended for teacher self-appraisal and was one of the first of its kind. Barr (1931) regards Rugg's rating scale as "elaborate" compared with other self-rating scales, many of which contained unorganized lists of questions. Barr offers several examples of diverse rating scales.

The "Boyce Card" (Boyce, 1915) received particular acclaim during the twenties and thirties, and was frequently cited as an example of a useful rating scale (see Figure 12.1). This card was also widely used for supervisory purposes, with minor modifications made from time to time. It was divided into five main sections: personal equipment, social and professional equipment, school management, technique of teaching, and results. These sections were further subdi-

vided, and teachers were rated as "excellent," "medium," "poor," or "very poor" on each of these subdivisions. The terms used in this scale are fairly precisely defined, an aspect of this rating instrument that is commendable and absent from most other rating devices of the period.

In addition, Barr (1931) describes a set of rating scales designed to be more objective than those that were in general use. These instruments, which he called

ANTIGO, WIS.

REPORT ON THE TEACHING EFFICIENCY

(To help teachers discover their own strength and weakness)

of

Teacher's name

City Building Grade taught

Years of experience Salary

Academic preparation

Professional preparation

Recorded by

Position

Date

Directions

1. Read the Explanation of Terms before making the following report:
2. Opposite each quality is a scale on which may be indicated all shades of judgment from "Very poor" to "Excellent." After due consideration place on this scale a small cross (x) indicating your judgment of the teacher as to that quality.

Example

Qualities of Merit	Very Poor	Poor	Medium	Good	Excellent
General Appearance.....				..X..	
Health.....		..X..			
Voice.....			..X..		

This card often contains from two to eight divisions in each of the five columns given, thus making possible shades of judgment.

3. Record your judgment on General Rating in the same way after completing the detailed record. General Rating is your estimate of general merit and is to be made on the basis of the detailed record.

Figure 12.1 The Boyce Card.

Explanation of Terms

- I. Personal equipment includes physical, mental and moral qualities.
1. General appearance: Physique, carriage, dress, and personal neatness.
 3. Voice: Pitch, quality, clearness of schoolroom voice.
 4. Intellectual capacity: Native mental ability.
 5. Initiative and self-reliance: Independence in originating and carrying out ideas.
 7. Accuracy: In statements, records, reports, and school work.
 10. Integrity and sincerity: Soundness of moral principles and genuineness of character.
 13. Tact: Adroitness; address; quick appreciation of the proper thing to do or say.
 14. Sense of justice: Fairmindedness; ability to give all a "square deal."
- II. Social and professional equipment includes qualities making the teacher better able to deal with social situations and particularly the school situation.
1. Academic preparation: School work other than professional.
 2. Professional preparation: Preparation of a technical nature.
 3. Grasp of subject matter: Command of the information to be taught or the skill to be developed.
 4. Understanding of children: Sympathetic insight into child nature of a scientific and practical kind.
 5. School and community interest: Interest in the life of the school and of the community.
 7. Interest in lives of pupils: Desire to know and help pupils personally, both as to their present condition and future prospects.
 8. Cooperation and loyalty: Attitude toward colleagues and superior officers.
 9. Professional interest and growth: Effort to keep up to date and improve.
 11. Use of English: Vocabulary, grammar, ease of expression.

- III. School management includes mechanical and routine factors.
3. Care of routine: Saving time and energy by reducing frequently recurring details to mechanical organization.
 4. Discipline (governing skill): Character of order maintained and skill shown in maintaining it.
- IV. Technique of teaching includes skill in actual teaching and in the conduct of the recitation.
1. Definiteness and clearness of aim: Of each lesson and of the work as a whole.
 2. Skill in habit formation: Skill in establishing automatic responses quickly and permanently.
 3. Skill in stimulating thought: Giving opportunity for and direction in reflective thinking.
 4. Skill in teaching how to study: Establishing economical and efficient habits of study.
 5. Skill in questioning: Character and distribution of questions; replies elicited.
 6. Choice of subject matter: Skill with which the teacher selects the material of instruction to suit the interests, abilities, and needs of the class.
 7. Organization of subject matter: The lesson plan and the system in which the subject matter is presented.
 9. Skill in motivating work: Arousing interest and giving pupils proper incentives for work.
 10. Attention to individual needs: Teacher's care for individual differences, peculiarities, and difficulties.
- V. Results include evidence of the success of the above conditions and skill.
1. Attention and response of the class: Extent to which all of the class are interested in the essential part of the lesson and respond to the demands made on them.
 2. Growth of pupils in subject matter: Shown by pupils' ability to do work of advanced class and to meet more successfully whatever tests are made of their school work.
 3. Social development of pupils: Increase in pupils' ability and power along lines other than those of subject matter.
 4. Stimulation of community: Effect on life of the

Figure 12.1 continued

Qualities of Merit	Very Poor	Poor	Medium	Good	Excellent
<i>I. Personal Equipment:</i>					
1. General appearance.....					
2. Health.....					
3. Voice.....					
4. Intellectual capacity.....					
5. Initiative and self-reliance.....					
6. Adaptability and resourcefulness.....					
7. Accuracy.....					
8. Industry.....					
9. Enthusiasm and optimism.....					
10. Integrity and sincerity.....					
11. Self-control.....					
12. Promptness.....					
13. Tact.....					
14. Sense of justice.....					
<i>II. Social and Professional Equipment:</i>					
1. Academic preparation.....					
2. Professional preparation.....					
3. Grasp of subject matter.....					
4. Understanding of children.....					
5. School and community interest.....					
6. Ability to meet and interest parents.....					
7. Interest in lives of pupils.....					
8. Cooperation and loyalty.....					
9. Professional interest and growth.....					
10. Daily preparation.....					
11. Use of English.....					

Qualities of Merit	Very Poor	Poor	Medium	Good	Excellent
<i>III. School Management:</i>					
1. Care of light, heat and ventilation.....					
2. Neatness of room.....					
3. Care of routine.....					
4. Discipline (governing skill).....					
<i>IV. Technique of Teaching:</i>					
1. Definiteness and clearness of aim.....					
2. Skill in habit formation.....					
3. Skill in stimulating thought.....					
4. Skill in teaching how to study.....					
5. Skill in questioning.....					
6. Choice of subject matter.....					
7. Organisation of subject matter.....					
8. Skill and care in assignment.....					
9. Skill in motivating work.....					
10. Attention to individual needs.....					
<i>V. Results:</i>					
1. Attention and response of the class.....					
2. Growth of pupils in subject matter.....					
3. Social development of pupils.....					
4. Stimulation of community.....					
5. Moral influence.....					

checklists, include general checklists, activities checklists, and lists of standard items to observe. General checklists, the least objective of the three, consist of lists of teacher characteristics that had been determined somewhat subjectively. A teacher-characteristic profile was provided by checking the characteristics considered to be present by the observer. Activity checklists were used to record all activities occurring in the classroom without regard to their value or appropriateness. These checklists could be quantitative (e.g., determining how much of a particular activity was occurring) or qualitative (e.g., for recording the presence or absence of certain teacher comments, the use of certain kinds of assignments, etc.). Standard items to observe, Barr's third category of checklist, were more objective than the general or activities checklists. They consisted of lists of specific teacher and pupil activities of which the observer noted the presence or absence. Barr himself worked for several years trying to produce more objective "items-to-observe" checklists.

Content. During the 1930s and 1940s several researchers compared the content of various rating devices. Morsh and Wilder's (1954) analysis of the findings of some of these studies led them to conclude that:

There is no general agreement as to what constitutes the essential characteristics of a competent teacher. Similarly, items on present rating scales tend to be subjective, undefined, and varied, there being no consistency as to what traits a supervisor might be expected to observe and evaluate (p. 3).

Most researchers in this area examined a large number of rating instruments, tabulated the items on these instruments, and determined the frequencies of occurrence of the various traits or characteristics. Findings from these surveys revealed a notable lack of uniformity among rating methods. There was variety in the number of items to be rated, the characteristics included, and the individual response for the rating (Barr & Emans, 1930).

In a frequently cited study of rating scales, Barr and Emans (1930) examined 209 rating scales collected from U.S. cities of more than 25,000 population. From these scales, 6,939 items were classified into 200 groups, with frequencies for each group indicated. A more compact, though relatively subjective classification was finally produced that included seven main categories as follows: (1) classroom management; (2) instructional skill; (3) personal fitness for teaching; (4) scholarship and professional preparation; (5) effort toward improvement; (6) interest in work; and (7) ability to cooperate with others. Barr and Emans found no major discrepancies between their findings and the results of the Commonwealth Teacher Training Study (Charters & Waples, 1929).

Morsh and Wilder (1954) are critical of the rating scale analysis studies. They acknowledge that these studies might be of some value in developing improved rating scales, but they consider that factors such as problems with categorization procedures render the results of these studies meaningless. Nevertheless, an examination of the content of rating scales is a useful means of determining what

variables were considered important enough by rating scale developers to be included in such instruments and the bases on which teachers were appraised and evaluated.

Use and consequences. Rating scales were used by a variety of individuals and groups who frequently tailored the scales to their own purposes. Teachers might have been rated by superintendents, supervisors (administrators primarily responsible for supervision and accountable to the superintendent or to one of his assistants), principals, researchers, teacher district-training personnel, other teachers, pupils, or by teachers themselves. Whereas administrators usually rated teachers, a survey reported by Barr and Burton (1926) reflected the emerging key role of principals in rating teachers. The survey showed that of 75 teachers who reported that they were rated, 41 were rated by supervisors with or without others, 65 were rated by principals with or without others, and 32 were rated by superintendents with or without others.

The supervisors' role was wide-ranging and diffuse, with direct supervision of teachers only one responsibility among many. A study carried out in Detroit (Barr & Burton, 1926, pp. 11-13) showed that appraisal was 1 of at least 13 tasks that supervisors were expected to perform. Appraisal included recommending teachers for appointment, rating teachers, advising administrators about the transfer and assignment of teachers, and assisting teachers with personal problems.

Rating by administrators for evaluative purposes was the first and most prevalent use of rating scales. Scales were used for promotion, transfer or dismissal of teachers, merit judgments, and public accountability. Rating of teachers for promotion, transfer, or dismissal was conducted largely by supervisors and principals who were considered most knowledgeable about teachers' work. Ratings were sometimes used as a basis for salary increases, but their use for this purpose was highly unpopular with teachers. Teacher opposition and administrative apprehension prevented the widespread use of this form of rating, which became known as *merit rating*. Administrative rating was also used to obtain information on teachers' proficiency in order to justify educational expenditures to the public. Students' standardized test scores were one source of information on the quality of schooling. Some measure of teacher efficiency was also needed, and rating scales were perceived as being useful for this purpose.

Barr and Burton (1926) made an important distinction between administrative and supervisory rating. In their view, supervisory rating differed considerably from administrative rating and had the following purposes: "(a) to improve the teacher in service; (b) to stimulate self-analysis, self-criticism, and self-improvement; (c) to raise professional standards; and (d) to provide supervisors and principals with uniform standards with which to judge teachers, analyze their work, and stimulate improvement" (p. 450). In 1931, Barr predicted that the supervisory uses of ratings would overtake administrative uses, although administrative rating would continue to be employed. He held that there was a danger of negative attitudes towards administrative rating being transferred to supervi-

sory rating, together with the creation of more objective and valid rating tools. The development and use of cooperative rating systems, and of self-rating and student rating of teachers from the late 1930s illustrates the beginning of a more enlightened perspective on supervision in the direction of Barr's recommendations.

Difficulties. At that time of Barr and Burton's writing in 1926, the main arguments advanced against rating were the following:

1. Rating is unprofessional; it prevents teaching from becoming a profession.
2. Rating differs with individuals and even with the same individual over time and is therefore manifestly unfair.
3. Rating is open to too many abuses, makes unfair discrimination very easy, and is therefore unsafe.
4. Rating schemes force the teacher to live up to the scheme and not to teach to good pedagogical principles. Rating forces the teacher to play to the rater and not to the children's interests" (p. 457).

These criticisms were directed toward administrative rating, and Barr and Burton believe that many of them are justified. They consider that using supervisory rating more widely could be an answer to the first criticism. If rating was conducted to improve teachers' in-service, the status of the profession might be enhanced. The creation of more objective rating devices is suggested as a means of combating the second criticism. Barr and Burton agree that rating systems can be open to abuse but that they can also support and mask inefficiency. They suggest that supervisory rating that involves teachers cooperating in their own improvement is likely to avoid the danger of the final criticism. The arguments advanced against rating in the early part of the century are similar in many ways to arguments advanced against teacher assessment and evaluation procedures today, and Barr and Burton's response to these arguments applies in many ways here also.

By the mid-1950s there was a large teacher effectiveness literature and several bibliographies in this area had been produced (e.g., Morsh & Wilder, 1954). This research, however, was in general poorly regarded. The criticism of early teaching effectiveness research is reflected in the following comment of the Committee on Criteria of Teacher Effectiveness of the American Educational Research Association (1953):

The simple fact of the matter is that, after 40 years of research on teacher effectiveness during which a vast number of studies have been carried out, one can point to few outcomes that a superintendent of schools can safely employ in certifying teachers, or that a teacher education faculty can employ in improving teacher education programs (p.65).

Medley and Mitzel (1963) argued that much work on teacher effectiveness should be disregarded as irrelevant because it lacked criteria and objective measures. Hence, since its inception, teacher rating has produced little certified knowledge about what should be rated and how rating can best be accomplished.

Historical Problems of Teacher Evaluation

In this section, the authors summarize important influences on teacher evaluation. It is argued that various problems prevented teacher evaluation practices from generating useful knowledge.

External Evaluation

Rating systems were developed primarily for reasons external to teachers, that is, to demonstrate to the public that students were receiving appropriate instruction or that teachers were competent, rather than to provide teachers with information that they might use to improve instruction. Teacher evaluation has usually been performed by administrators who typically developed the criteria and procedures for assessing teachers without consulting teachers, teacher educators, or citizens.

Complexity of Teaching

Teaching has been consistently viewed as a multidimensional task. Teachers were seen as planning, providing leadership, developing curriculum, and delivering and evaluating student learning. Likewise, those who developed rating systems generally did not view teaching in simple terms. Indeed, evaluation checklists of desirable behaviors and personality characteristics, if anything, tended to contain an overwhelming number of factors to be rated.

Lack of Attention to Context or Theory

Despite administrators' awareness of the complexity of teaching, there is little evidence that they considered that different types of teaching were required in various classroom and school contexts. There was not much concern for the effects of grade, subject matter, types of learners, or instructional goals on teacher effectiveness. Further, often there was no theory or framework for ordering the factors involved in teaching.

Minimum of Observation

Administrators believed that they could rate teachers on many of the behaviors or personality characteristics on checklists with little or no classroom obser-

vation (perhaps on the basis of teachers' conversations in the teachers' lounge). Questions were sufficiently general that not much observation was required to complete the ratings. As noted earlier, most ratings were done by administrators; however, occasionally teachers or students evaluated teachers. Even when teachers were asked to rate other teachers, they also typically did so after little or no classroom observation.

The Contemporary Problem of Teacher Evaluation

It is instructive that despite years of concern over issues of teacher evaluation, the problems described above continue to plague evaluation. There is *no* clear evidence that teachers or administrators learn from evaluation. The examination of teacher evaluation practices has led some researchers to conclude that few effective teacher evaluation systems exist in the United States (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1985). Indeed, some critics decry the entire enterprise as a failure. According to Scriven (1981), "teacher evaluation is a disaster. The practices are shoddy, and the principles are unclear" (p. 244).

Why, despite the long history of teacher assessment, is there such widespread disagreement about the *purposes* and procedures for teacher evaluation? There are at least six reasons for the problem. First, there is public ambivalence about the professional role of teachers. Second, there is no serious investment of research funds to understand the evaluation process. Third, school districts often have a variety of evaluation goals and procedures and tend to use the same procedures to pursue disparate goals. Fourth, too little is known about the relations between teacher actions, student classroom behaviors, and various learning opportunities and specific student outcomes. Fifth, because the knowledge of teaching is limited, school districts tend either to ignore research or to rely excessively on research results. Sixth, the evaluation process often becomes a ritual that principals and teacher engage in because it is expected — not because they value it. Having discussed possible reasons for disagreement about purposes and procedures of evaluation, the authors now turn to a discussion of five major *problems* with extant evaluation practices.

External Pressures for Evaluation

The National Commission on Excellence in Education's 1983 report, *A Nation at Risk*, resulted in an outpouring of public concern over schooling that continues in today's political discussion. For example, on the day we began this chapter, the following headline was emblazoned on the front page of the June 4, 1988, edition of the Chicago *Sun-Times*: "**BUSINESS OKS TAX BOOST — IF SCHOOLS REFORM.**" Among the reforms called for in that article were re-training mediocre teachers, providing principals with more power to control

their schools, and giving districts more autonomy to judge the value of curriculum. It is interesting that the reforms called for assumed that teachers were the problem and that principals had the necessary expertise to improve schools.

Public dissemination of information about the problems of schooling has been extensive. Tomlinson (1986) notes that the full text of the report, *A Nation at Risk*, has been placed into the hands of at least four million citizens (using conservative calculations). Within months after the report's initial release, the *New York Times* alone had published about 50 articles dealing with it.

An interesting aspect of the new reform movement is that much of the rating of teachers and schools is done through publishing student scores on standardized achievement tests in public newspapers. Not content with publishing the test scores, some policy makers have also used newspapers and other public forums to describe the effectiveness of school districts. For example, on Thursday, May 12, 1988 the following bold headline appeared in the *St. Louis Post Dispatch* newspaper: "BENNETT CALLS ST. LOUIS SCHOOLS 'AWFUL'." At a meeting on Wednesday, May 11, 1988 in Springfield, Illinois, Bennett said that the St. Louis Public School System might even be worse than the Chicago system, which he had already labeled as the worst in the nation. "If you look at St. Louis, you'll probably find a higher expenditure per student and maybe even lower (student) performance than Chicago," Bennett added, "It's awful."

Policy makers' comments have caused citizens to worry about how schools in their state compare to schools in other states. The public has been made keenly aware of the poor academic performance of American youth compared to Japanese or Australian youth. However, there are some growing signs that the public may have learned of the inadequacies of these tests as an index of the quality of schooling in a particular district, school, or classroom.

It is significant that on the same day that Chicago schools were being offered money if they would reform, an editorial in the *St. Louis Post-Dispatch* described dissatisfaction with how the outcome of reform was assessed:

"... school leaders such as Education Secretary William Bennett began clamoring for concrete measurements that would show taxpayers and students were getting their money's worth in the nation's classrooms; the best answer they have come up with so far is more standardized tests.

That answer is wrong, for many reasons. When too much stress is put on test scores — to the point that in St. Louis, teachers' evaluations depended on them — the temptation to cheat or teach directly to the test can be overwhelming, particularly when the same version of a test is used year after year. Students end up learning primarily what is on the test, regardless of whether that is the best course of study. They also are shortchanged in the development of the kind of critical thinking and problem-solving skills that are essential for true learning but are not easily tested.

Standardized tests need not be banned, but their use should be reduced and put into the proper perspective. Not everybody needs to learn the same things in the

same grade. Curriculum should be controlled locally, and schools whose courses are a little different should not have to suffer by comparison. Like all tools, tests can be used well or poorly. They are not designed to fix everything, and in the wrong hands, they can do more harm than good" (June 4, 1988).

Public Ambivalence About Teachers

At the time the authors were revising this chapter, the following two headlines appeared on the same page of the *Columbia Daily Tribune* (September 11, 1988): (a) "TEACHERS SAY THEY'RE SNUBBED"; (b) "STUDY SAYS TEACHERS IGNORANT." The first headline summarized results of a survey of 21,698 public school teachers by the Carnegie Foundation for the Advancement of Teaching. Teachers reported that they had only a limited part in important decisions such as those concerning teacher evaluation or selection of new teachers or administrators. The second headline was associated with a national survey released by Raymond Eve and Dana Dunn, two university of Texas-Arlington researchers who contend that a high percentage of high school biology teachers profess superstitious beliefs. Headlines such as this one have become common in contemporary publishing. Hardly a week goes by without teachers being judged in newspapers.

Teachers are also judged by national surveys. For example, one influential journal, the *KAPPAN* (Phi Delta Kappa), conducts an annual survey of public opinion polls concerning the efficacy of teachers and schooling. Further, there is the constant informal judgment of teachers on dimensions other than standardized measures of student achievement. Informal communication, such as students' comments to parents or parents' comments to other parents, may have as much or more of an effect on teachers' informal status than do more formal measures (teachers' ratings on classroom process scales or measures of student achievement).

Although citizens are extremely concerned about schooling, they also want to make the work-place conditions of teaching more attractive and the profession more rewarding. The results of various political pressures to improve schools remain problematic, however. It is safe to say that some of the longstanding conditions in education (e.g., external accountability) are at least open to examination and modification.

Inadequate Research on Teacher Evaluation

Surprisingly, despite the widespread practice of teacher evaluation, there is little evidence that teachers, administrators, or the public learn much from evaluation. In a wide-scale study of teacher evaluation, Wise et al. (1985) concluded that the way in which evaluation is currently conducted allows only a few districts to identify excellent teachers. Moreover, they contend that even

fewer districts can identify — on reliable and valid grounds — teachers who are not minimally competent.

Wise et al. did find a small number of districts that identified inadequate teachers and improved their performance. These districts had no quick-fix procedures. Rather, their success was achieved through careful planning, administrative time, money, and systematic implementation of an evaluative process (i.e., training of evaluators, etc.). Thus, the quality of evaluation was more important than its form. The extent to which evaluators (teachers or administrators) possessed wide knowledge of teaching and gave ample time to the task was more important than the procedures or the rating form utilized.

These researchers argue that, because of limited resources, school districts need to set realistic, specific goals for evaluating teachers. As Wise et al. note, a system that identifies inadequate teachers will not be successful in identifying excellent teachers. Similarly, an evaluation system that rates teachers for merit awards will not be ideally suited for developing plans to improve teachers' skills.

Failure to Match Evaluation and Context

A major issue in the appropriateness of a rating device is the type of decision it will be used to make (merit pay, tenure, improvement of instruction). It may be appropriate to assess how well beginning teachers can structure classrooms for learning, teach basic concepts, or offer effective demonstrations. However, it makes no sense to assess more experienced teachers on these criteria or to base career ladder programs on them (Wise et al., 1985).

Many important aspects of teaching remain unstudied. For example, the curriculum demise depicted in *The Shopping Mall High School* and the aimlessness and lack of belonging (associated with a large mall) suggest that teachers must know how to design intellectually stimulating curricula and classrooms that offer a sense of community if they are to be effective (see Powell, Farrar, & Cohen, 1985). Yet building a sense of community may well entail different procedures and values in an inner-city Chicago high school than in a suburban elementary school. Too much emphasis has been placed on specific teacher behaviors and too little on student proxies (attendance, ability to ask questions, interest in school, willingness to take academic courses) that could signify progress in a particular school. Surely school and teacher effectiveness might appear radically different in diverse settings (Getzels & Jackson, 1963). However, despite long recognition of the problem, evaluation systems that do not take context into account are still in use.

Confusion About the Objectives of Evaluation

There are many reasons to evaluate teachers. One may wish to assess the effects of teaching on students' achievement; however, one could focus on various types of achievement (e.g., basic skills, problem solving). Teacher eval-

uation may be performed only for symbolic reasons (e.g., it is required by law), and neither veteran teachers nor administrators may expect much from the process. In other cases, important personnel decisions may rest on evaluation.

Part of the problem is that participants are often unclear about the objectives of evaluation. Evaluators must understand not only what evaluation process is appropriate in a given context but also the goals of evaluation. Further, when clarity is not the issue, districts often do not realize that different evaluation objectives need to be addressed with various types of instruments and processes. Darling-Hammond (1986) notes that in too many cases all teachers are evaluated using one instrument that simultaneously serves as the primary means of rating teaching practice and professional development as well as the basis for personnel decisions.

Process-Product Research: A Source for Teacher Ratings

The quest for information about the effects of various teaching behaviors led to process-product research in the 1970s and early 1980s. One useful result of this research was evaluation relating certain classroom processes to students' scores on standardized achievement tests as well as on criterion-referenced tests linked directly to the goals of classroom instruction. Considering the hundreds of possible relations between process and product, this was a small but significant step (see Brophy and Good, 1986, for a review of this literature).

Unfortunately, many practitioners and policy makers have interpreted results of process-product research to mean that competent teachers only need good communication skills. However, successful teaching involves more than effective communication (appropriate pace, clarity, good use of examples); it also entails the application of principles of human cognitive development; an understanding of human motivation; knowledge of subject matter; as well as pedagogical skills necessary to help students to understand curriculum (Doyle, 1988; Shulman, 1986). As Griffin (1986) notes, it is encouraging that the knowledge base for teaching is expanding, but it is important to recognize what research has not examined and to expand this knowledge base in subsequent research.

Knowledge from process-product research. The significance of process-product research is that it yields useful instructional models that are relevant to instruction in lower-level content and skills. These models can also be applied to some higher-level activities such as reading comprehension and study skills and certain types of scientific experimentation and mathematical problem solving (e.g., principles that can be presented systematically). However, process-products research does not provide information about instruction involving processes rather than products (debate, discovery, discussion, role play, simulation) or tasks that call for writing to discover or create processes (interpretation of literature, creative writing, etc.) rather than to follow prescribed processes.

The overgeneralization of research findings. Although information from research is important, some policy makers (and accountability systems) have used it to distort the image of teaching. They fail to recognize the limitations of findings from process-product research and tend to equate limited but important knowledge about one aspect of teaching with effective teaching in general. They often forget that basic skills instruction must be adapted to subject matter, students, and other classroom context factors. Thus, research linking teacher behavior to student outcomes does not yield rules or simple answers to complex issues of teaching and learning (Good & Brophy, 1987; Shulman, 1986; Zumwalt, 1986).

As we noted, one advantage of the rating of teachers is that it has always been based on a relatively broad definition of a good teacher. Any list of characteristics of such a teacher should be considerably broader than the knowledge base for relating instruction to student achievement. For example, Raths (1971) suggested the following framework for teachers to use as they consider their performance as classroom teachers. According to Raths, teaching involves at least the following functions: (1) explaining, informing, showing how; (2) initiating, directing, administering; (3) unifying the group; (4) giving security; (5) clarifying attitudes, beliefs, problems; (6) identifying learning problems; (7) making curriculum materials; (8) evaluating, recording, reporting; (9) enriching community activities; (10) organizing and arranging classrooms; (11) participating in school activities; and (12) participating in professional and civic life.

Research provides data relevant to only three of Raths' teaching functions (1, 2, and 10), and these findings apply only in certain contexts. That is, the data concerning explaining, informing, and showing how are related to knowing about how teachers can give successful demonstrations and teach concepts in mathematics and reading. Missing are explanation of how complex problem solving and higher-order thinking might be successfully taught.

Still, extant research is useful when it is applied appropriately. Teachers must convey information in a way that promotes student thinking and comprehension. Knowledge about how to do this successfully is important, especially as there is considerable evidence that a number of teachers have difficulty managing classrooms and promoting students' comprehension of basic concepts. Despite the contributions of research, it would be foolish to equate knowledge of such teaching effects with effective teaching (Brophy & Good, 1986).

Overgeneralizing the utility of knowledge from process-product research on teaching is similar to equating a good physician with the ability to interpret an X-ray correctly, defining a successful mechanic as one who can tune a car, or reducing a competent basketball player to one who can hit a 15-foot jump shot. The skills and knowledge required to be successful in any vocation are numerous and diverse (a good physician must understand cell structure, bone structure, as well as principals of physiology and pharmacology, to name but a few; a good mechanic must know about carburetors, fuel-injection systems, principles of suspension, and how to identify the cause of engine malfunctions; or a basketball

player must have skills of defense, rebounding, running the court, and passing, as well as court presence).

It is unfortunate that in some state accountability systems one subpart or subskill of teaching (how to present demonstrations and information in an interesting and comprehensible fashion) is often equated with good teaching in general. This silly policy has angered some individuals to the point that they reject information about these aspects of teaching as unimportant or unprofessional. This decision is equally inappropriate, because this knowledge from research does apply in certain contexts.²

Unfortunately, the appropriate use of research on teacher effects has not been observed in many situations. Brophy (in press) summarizes this misuse: "Research on teacher effects has been seriously misused in many teacher evaluation and accountability programs developed by state departments of education and local school districts. Any such effort that in effect imposes a single lesson format on all teachers in all teaching situations is simply invalid, and cannot be justified by claiming that it is supported by research on teacher effects." We concur with this judgment.

The Future Role of Ratings

It should be clear that teacher evaluation has not yielded unambiguous criteria for evaluating teachers. Some progress has been made, and further research will eventually yield information about critical areas that are poorly understood (How can teachers help students to acquire skills and knowledge relevant for problem solving and self-evaluation?). However, knowledge of effective teaching process will always be indeterminant, and we will always want teachers to be concerned about student outcomes that are difficult to achieve.

Although we provide some examples of rating forms in this chapter, it would be presumptuous to advocate the use of certain forms. If we have learned anything from studies of practice, it is that school systems that learn from evaluation carefully match their objectives with the evaluation procedures they use. Further, if evaluation has multiple purposes, then several procedures are required. Finally, evaluators' knowledge (research on teaching effects, child development, motivation, subject matter) and ability to use knowledge are much more important than the particular rating scale utilized. The authors are not suggesting that factors such as training and reliability are not important nor that all classroom dimensions deserve equal emphasis in ratings. No matter how competent evaluators are, they need to demonstrate that they apply a system fairly and consistently.

The authors also recognize that the administrator's role requires that administrators make personnel decisions about teachers. Rating systems can be useful in helping to make certain personnel decisions. However, considering knowledge about teaching and the inadequate funds for training people to use rating sys-

terms, career decisions will still be somewhat subjective. Thus, those who use rating systems to make personnel decisions must do so conservatively and must supplement them with other evidence. The authors encourage the inclusion of teachers in this process.

In the authors' judgment, the key role for teacher ratings in the 1980s is to expand opportunities for teachers to reflect on instruction by analytically examining classroom processes. For too long rating systems (classroom observation systems, and related evaluative procedures) delineated what teachers should do and collected information about the extent to which they did it. Ratings of teacher behavior should be made not only to confirm the presence or absence of a behavior but with the recognition that many aspects of a teaching behavior are important (quality, timing, and context), and that numerous teacher behaviors combine to affect student learning.

Work-Place Conditions of Teaching

There is growing recognition that the work-place conditions of teaching are often difficult and unrewarding. As the Holmes Group Report, *Tomorrow's Teachers* (1986) notes, teachers must instruct classes all day long with little or no time for preparation, analysis, evaluation of their work, or opportunity to exchange intellectual ideas with colleagues. According to Freedman, Jackson, and Boles (1983), teachers work in an institution that has the putative goal of preparing students for adulthood but which too often views those responsible for this socialization task — the teachers themselves — as incapable of mature judgment. Although teachers are often recognized as an important part of better schools, they are also frequently the object of criticism and defined as a major part of the problem. Hence, the combination of a workplace that allows little opportunity for reflection but demands a vast amount of physical energy, and frequent (and often unjustified) external evaluation in the form of newspaper headlines and other public comment is not conducive to constructive teacher dialogue and school reform.

Making Teaching a Profession

Growing teacher dissatisfaction is having a serious, negative effect on the attractiveness of teaching as a profession (talented youth are not considering teaching as a career; capable teachers leave the profession after only one or two years). This has led many educators to argue that until teaching becomes more of a profession, it is unreasonable to expect the field to attract a sufficient number of talented individuals to public schools (Schlechty & Vance, 1983).

One proposed change involves shifting from external evaluation to evaluation that is partially controlled by teachers in their own setting. Darling-Hammond (1987-88) argues that to make teaching more professional it will be necessary to

give teachers more authority and responsibility over curriculum and assessment of student mastery of important, fundamental knowledge.

*The Future Role of Teacher "Rating"
and "Descriptive Feedback" Systems*

Historical analysis of the use of teacher rating systems, as well as of contemporary issues that confront the teaching profession, has led the authors to conclude that teachers can use evaluation systems to study classroom behavior and to analyze and interpret information obtained from such study. Because of the multiple purposes of teaching and the fact that career teachers — even at the same grade level in the same school — can teach in different classroom contexts, descriptive scales are useful because they provide teachers with rich, detailed information from a variety of sources (self, peers, students, administrators).

In this section of the chapter some uses of descriptive scales are explored. The authors believe that teachers should have more responsibility in these areas so that observations provide supervisory and personal information that is relevant to individual teachers.³ By *descriptive scale*, the authors imply that the intent is largely to describe what did or did not occur and to share inferences about the quality of the behavior, curriculum tasks, climate, or whatever is being described. The intent is to stimulate a conversation — an analysis — not to render a final judgment as may be the case in administrative ratings.

**Dealing with Classroom Complexity:
Descriptive Scales**

There are simply too many aspects of classroom instruction to allow the development of a single, systematic observational scale that could comprehensively assess teaching. The authors believe that teachers will benefit from and welcome information that describes teaching in a way that allows them to reflect on it. That is, descriptive systems could stimulate thinking about teaching in a given school context and would include many descriptive scales capable of providing information across a variety of classroom dimensions. The scales utilized on a given day would be related to the instructional format and emphasis of instruction and/or teachers' preferences (e.g., interest in one aspect of classroom behavior).

*Flexible Descriptive Scales:
An Example*

Desirable instruction varies from classroom to classroom, and the descriptive scale used should vary accordingly. Descriptive rating systems possess the

FORM 8.3. General Motivational Strategies

USE: When the teacher has been observed frequently enough so that reliable information is available

PURPOSE: To assess the degree to which the teacher's general approach to instruction supports students' self-confidence and motivation to learn

How regularly does the teacher follow the motivational guidelines listed below? Rate according to the following scale:

- 5 = always
- 4 = most of the time
- 3 = sometimes
- 2 = occasionally
- 1 = never

A. ESSENTIAL PRECONDITIONS

- _____ 1. Maintains a supportive learning environment (classroom atmosphere is businesslike but relaxed, teacher supports and encourages students' learning efforts)
- _____ 2. Assigns tasks at an appropriate level of difficulty (students can achieve success with reasonable effort)
- _____ 3. Assigns tasks with meaningful learning objectives (the tasks teach some knowledge or skill that is worth learning)
- _____ 4. Demonstrates moderation and variation in use of motivational strategies (does not overuse particular strategies to the point that they become counterproductive)

B. MAINTAINING STUDENTS' SUCCESS EXPECTATIONS

- _____ 1. Programs for success
- _____ 2. Helps students to develop their skills for goal setting, performance appraisal, and self-reinforcement
- _____ 3. Helps students to recognize effort-outcome linkages
- _____ 4. Portrays effort as investment rather than risk
- _____ 5. Portrays skill development as incremental and domain-specific (rather than as determined by fixed general abilities)
- _____ 6. Focuses on mastery in monitoring performance and giving feedback (stresses student's continuous progress toward mastery rather than comparisons with other students)
- _____ 7. If necessary, does remedial motivational work with discouraged students (to help them see that they have ability and can reach goals if they put forth reasonable effort)

C. STIMULATING STUDENTS' MOTIVATION TO LEARN

- _____ 1. Models own motivation to learn (portrays learning as a self-actualizing activity that produces personal satisfaction and enriches one's life)
- _____ 2. Communicates desirable expectations and attributions (implying that students see classroom activities as worthwhile and are eager to acquire knowledge and master skills)
- _____ 3. Minimizes performance anxiety (treats mistakes as understandable and expected; minimizes the threat of tests)

NOTES:

Figure 12.2 General Motivational Strategies.

This form is taken from Good, T. and Brophy, J. *Looking in Classrooms*, Fourth Edition (1987). Harper & Row, New York

advantage of adjusting what is examined to the interests and concerns of individual teachers.

To illustrate the flexibility of descriptive scales, consider the form that appears in Figure 12.2. The scale in Figure 12.2 is taken from the book, *Looking in Classrooms*, and would provide useful information for teachers who want to foster students' intrinsic motivation. Teachers who emphasize behavior modification would find little of value in the scale. Good & Brophy (1987) present over 30 scales in the book *Looking in Classrooms* that allow descriptions of various teacher-learning conditions.

Teacher Development of Descriptive Scales

There is no descriptive scale that is most appropriate. What is appropriate or important depends on the instructional goals and the reason that a teacher wants information. Although various sources are available for teachers to use in selecting scales, teachers can construct their own scales. For example, a group of third- and fourth-grade teachers who are beginning to use work groups in mathematics (small heterogeneous groups of students who work cooperatively on problem-solving tasks) might identify specific processes that they want to observe when students are interacting. Such attention to process might help teachers to become more sensitive to subtle problems associated with group work and to become better observers of groups. Teachers who become adept at instructing and/or using descriptive scales are also developing skills for action research. Teachers who can specify what they want to observe in terms of curriculum goals as well as instructional activities can collect more relevant evidence to determine which classroom procedures or activities (e.g., having students write their own accounts of certain historical events) result in more student interest in content (e.g., student questions during subsequent class discussion). As various educators have noted, there is too little classroom observational research (e.g., Good & Biddle, 1987). However, there is no reason why teachers cannot do classroom research themselves and certainly no reason why they should not have the necessary resources to do so in their own schools (e.g., secretarial support; release time to observe their own students while someone else teaches; or the opportunity to observe other teachers).

Teacher Study Groups

Good and Brophy (1987) argue the advantage of teacher study groups in which teachers share information and resources and analyze teaching in collegial settings. The authors believe that teachers can often use regularly scheduled in-service time or release time during the school day to plan and work in study groups. This procedure can be especially beneficial for small groups of teacher who share common problems or interests. Within such groups, teachers could develop descriptive scalings for peer observation. However, a critical prerequi-

site is that teachers identify information and issues about which they want descriptive information.

Administrator, Student, and Parent Ratings

In addition to broadening the scope of classroom activities evaluated, teachers can also expand the sources from which they obtain feedback. Teachers can collect descriptive information from administrators, students, and parents and use it to improve teaching. Administrators can observe many classrooms and share information based on these observations. Too often teachers think of administrators only as sources of evaluation; however, many administrators can provide information about teaching that can help teachers improve instruction.

Students (especially older ones) can provide rich feedback to teachers about their learning and reactions to particular teaching strategies (e.g., How long did the homework take to complete? Was there a need to call classmates for clarifications?). Teachers can use students' opinions and preferences to supplement information from tests and project work. Evidence suggests that students can make reliable questionnaire responses about classroom instruction (e.g., Manatt, 1987).

Similarly, parents could be asked to give their opinions about various classroom activities and issues, for example, their reactions to the use of student portfolios (i.e., attempting to show parents samples of students' academic work) or other marking devices besides report cards. Some parents might find the new grading system confusing or see it as arbitrary. Information from parents might identify areas of student performance or curriculum content about which they want information.

New Opportunities for Improvement

Teachers could collaborate to develop rating scales by which administrators, students, and/or parents could provide information. Because individual teachers may not have all the skills required to develop a student portfolio (or to develop new curriculum in a particular area) — and because of the interest in linking classroom experiences across grade levels — teachers may want to work with colleagues. Similarly, teachers could invite university professors or teachers in other districts (who have developed comparable programs) to work with them to develop descriptive scales.

Conclusion

Analysis of early teacher effectiveness research reveals the considerable and varied efforts in the first half of the twentieth century to understand and describe the factors contributing to effective teaching. Unfortunately, the focus was primarily on teacher evaluation for administrative purposes, and the instruments

used, especially the rating scales, were largely unsuited to this role. As Barr and Burton (1926) pointed out, criticisms of early teacher appraisal were mainly directed at administrative rating, and similar criticism of research in this area occurred at least up to the late 1950s. The rating instruments developed and the knowledge obtained from their use might have been better employed in helping teachers to analyze and improve their instruction. If this had been a significant focus of early teacher effectiveness research, it is unlikely that subsequent evaluations of the research would have been so devastatingly critical.

More recently, rating systems for assessing teachers generally (systems that make global judgments about many aspects of teaching) were replaced with instruments that provided detailed information about specific aspects of classroom teaching. The new wave of observational instruments was successful in relating some teacher behaviors (communication of concepts; classroom management) to students' classroom performance. However, attempts to transform this knowledge into rating forms were largely unsuccessful because knowledge about behaviors were overgeneralized and equated with effective teaching *per se*.

Considering the multidimensional nature of teaching and the importance of context, the authors believe that rating scales have a valuable role to play in teacher supervision and self-development. Rating forms are called *descriptive feedback scales* because the term *ratings* suggests a judgment or an immediate decision. In contrast, *descriptive scales* implies that obtained information is data that a teacher can use to form a judgment. Although the descriptive feedback scale may be used on occasion to obtain ratings as well as information from others, a teacher does not use that information — on one scale or from one viewpoint — in any final sense.

The value of descriptive feedback scales lies in their ability to provide representative feedback from a number of individuals about various topics. The goal of collecting such information is to stimulate reflection and discussion about teaching among teachers.

Notes

1. Howard Ebmeier (personal communication) suggests that some teachers are unable to carry out independent critiques of their teaching and that until teachers have exhibited a minimal competency for teaching (i.e., have completed a successful administrative and peer review and been granted tenure) they should not be a part of a "self-development" group. The authors would agree that it makes sense for teachers to satisfy a minimum set of criteria that have been established by administrators and peer teachers. However, in this chapter, the authors do not focus on the marginal teacher nor the special needs of beginning teachers, but rather on the needs of teachers who have experienced a satisfactory adjustment to teaching.

2. Professor Richard Manatt (personal communication) reminds us that everyone enjoys teacher-performance-evaluation "bashing" but that without it, it is arguable that we might not have had the latest round of pay increases for teachers. Further, he notes that professors can debate the merits of teacher performance evaluation; however, school officials are required by law to do it!

3. Professor Richard Manatt, who has long been active in the area of teacher evaluation, appropriately suggests that administrators, as well as peers, can be helpful in improving performance. Some of the areas he suggests (personal communication) are written agreements, coaching, personalized staff development, and assistance plans for weak teachers.

References

- American Education Research Association, Committee on the Criteria of Teacher Effectiveness (1953). *Second Report of the Journal of Educational Research*, 46, 641-658.
- Anderson, W. N. (1917). The selection of teachers. *Educational Administration and Supervision*, 3, 83-90.
- Barr, A. S. (1931). *An introduction to the scientific study of classroom study of classroom supervision*. New York: Appleton Century.
- Barr, A. S., & Burton, W. H. (1926). *The supervision of instruction*. New York: D. Appleton.
- Barr, A. S., & Emans, L. M. (1930). What qualities are prerequisite to success in teaching? *Nation's Schools*, 6, 60-64.
- Baxter, B. (1941). *Teacher-pupil relationships*. New York: Macmillan.
- Beecher, D. E. (1949). *The evaluation of teaching: Backgrounds and concepts*. New York: Syracuse University Press.
- Boyce, A. C. (1915). Methods of measuring teachers' efficiency. *The Fourteenth Yearbook of the National Society for the Study of Education* (Part II). Chicago: University of Chicago Press.
- Brophy, J. (in press). Research on teacher effects: Uses and abuses. *Elementary School Journal*.
- Brophy, J., & Everson, C. (1976). *Learning from teaching: A developmental perspective*. Boston: Allyn & Bacon.
- Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.), (pp. 328-375).
- Charters, W. W., & Waples, D. (192). *The commonwealth teacher training study*. Chicago: University of Chicago Press.
- Darling-Hammond, L. (1986). A proposal for evaluating in the teaching profession. *Elementary School Journal*, 86, 531-569.
- Darling-Hammond, L. (1987-88). Between a rock and a hard place: The humanities and school reform. *Basic Education: Issues, Answers & Facts*, 3(2).
- Doyle, W. (April, 1988). *Curriculum in teacher education*. A paper presented at the Annual Meeting of the American Educational Research Association. New Orleans.
- Dunkin, M. J., & Biddle, B. J. (1974). *The study of teaching*. New York: Holt, Rinehart and Winston.
- Elliott, E. C. (1915). A tentative scale for the measurement of teaching efficiency. *Teachers Yearbook of Educational Investigators*. New York: Department of Research, New York City Schools.
- Freedman, S., Jackson, J., & Boles, K. (1983). Teaching: An imperilled "profession." In L. Shulman & G. Sykes (Eds.), *Handbook of teaching and policy*, (pp. 261-299). New York: Longman, Inc.
- Getzels, J. & Jackson, P. (1963). The teacher's personality and characteristics. In N. Gage (Ed.), *Handbook of research on teaching*, (pp. 506-582). Chicago: Rand McNally.
- Good, T., & Biddle, B. (1987). Research and the improvement of mathematics instruction: The need for observational resources. In D. Grouws & T. Cooney (Eds.), *Perspectives on research on effective mathematics teaching*. Hillsdale, NJ: Erlbaum.
- Good, T., & Brophy, J. (1987). *Looking in classrooms* (4th ed.). New York: Harper & Row.
- Griffin, G. (1986). *Teaching as a profession: The reform agenda for teacher education*. The Thirteenth Annual Robert Finkelstein Memorial Lecture, October, University of Rhode Island.
- Holmes Group, The (1986). *Tomorrow's teachers*. East Lansing, MI: The Holmes Group.
- Kratz, H. E. (1896). Characteristics of the best teachers as recognized by children. *Pedagogical Seminary*, 3, 413-418.
- Littler, S. (1914). Causes of failure among elementary school teachers. *School and Home Education*, 33, 255-256.
- Manatt, R. (1987). *Student feedback to teachers questionnaire*. Ames, IA: College of Education, Iowa State University.

- Medley, D. M. (1972). Early history of research on teacher behavior. *International Review of Education*, 18, 430-439.
- Medley, D., & Mitzel, H. (1963). Measuring classroom behavior by systematic observation. In N. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.
- Millman, J. (1981). Introduction in J. Millman (Ed.), *Handbook of teacher evaluation*, (pp. 12-13). Beverly Hills, CA: Sage Publications.
- Morsh, J. E., & Wilder, E. W. (1954). *Identifying the effective instructor: A review of the qualitative studies, 1900-1952*. Research Bulletin No. AFPTRC-TR-54-44. San Antonio, TX: USAF Personnel Training Research Center.
- National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Press.
- Ohio Teaching Record (1941). Ohio State University. (See Beecher, 1949)
- Powell, A., Farrar, E., & Cohen, D. (1985). *The shopping mall high school: Winners and losers in the educational marketplace*. Boston: Houghton-Mifflin.
- Raths, L. (1971). What is a good teacher. In J. Raths, J. Panchella, & J. Van Ness (Eds.), *Studying teaching*, (2nd ed.) (pp. 3-9). Englewood Cliffs, NJ: Prentice-Hall.
- Rugg, H. O. (1920). Self-improvement of teachers through self-rating: A new scale for rating teachers' efficiency. *Elementary School Journal*, 20, 670-684.
- Schlechty, P., & Vance, B. (1983). Recruitment, selection, and retention: The shape of the teaching force. *Elementary School Journal*, 83, 469-487.
- Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook on teacher evaluation*, (pp. 244-271). Beverly Hills, CA: Sage Publications.
- Shulman, L. (1986). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.), (pp. 3-36).
- Tomlinson, T. (1986). A nation at risk: Background for a working paper. In T. Tomlinson and H. Walberg (Eds.), *Academic work and educational excellence: Raising student productivity*, (pp. 3-28). National Society for the Study of Education: McCutchan.
- Troyer, M. E., & Pace, C. R. (1945). *Evaluation in teacher education*. Washington, DC: American Council of Education.
- Wise, A., Darling-Hammond, L., McLaughlin, M., & Bernstein, H. (1985). Teacher evaluation: A study of effective practices. *Elementary School Journal*, 86, 61-121.
- Zumwalt, K. (Ed.) (1986). *Improving teaching*. Alexandria, VA: ASCD Yearbook, 1986.

13. Self-Assessment

Larry W. Barber

Historically, self-assessment has been considered to be of little value. Most dismiss it as a strategy of self-improvement that is fraught with problems. Yet the numerous techniques of self-assessment can provide a powerful tool in improving the quality of teaching constantly sought by educators. To understand the utility of self-assessment techniques, one must view them in the context of formative evaluation.

What Is Formative Teacher Evaluation?

Formative teacher evaluation is a set of procedures designed to assist teachers in improving their own teaching. Formative teacher evaluation can help an ineffective teacher become a better teacher or an effective teacher become an excellent teacher. Formative evaluation is a helping, caring process that provides data to teachers for making decisions about how they can best improve their own teaching techniques, styles, or strategies. Formative evaluation must occur in close collaboration with the person being evaluated — he or she must agree to it, be an intensive part of it, participate willingly in it, and, in the case of experienced teachers, even direct it, thus a new dimension of self-assessment.

To improve teaching through formative evaluation, a teacher must first admit that he or she is doing something less than perfectly and that the teacher's behavior can be improved by one or a combination of the many techniques of formative evaluation. Such an admission requires much trust between the evaluator and the teacher being evaluated. Clearly, the formative evaluator must have the best interests of the teacher at heart before the teacher will admit that improvement is needed or possible. If these conditions are not met, then self-assessment techniques may be the only way that a teacher may be able to improve his or her own teaching.

AUTHOR'S NOTE: A great deal of assistance in developing this chapter was provided to the author by Phi Delta Kappa, Ernest Mueller, Roy Forbes, Julian Prince, Newman Walker, Sarah Lewis, Tom Saterfiel, and Phil Harris.

Where a trusting formative relationship exists, formative evaluators find that their role requires a commitment to the improvement of the teacher they assist. Evaluators become deeply involved with the professional lives of the teacher being evaluated. By taking on an assisting, helping, caring, supportive role, the formative evaluator becomes a major part of the teacher's professional development and almost always becomes ego-involved in seeing the teacher improve and become more productive in the classroom.

Formative evaluation provides a mirror to those being evaluated so they can see how to become better teachers. Formative evaluators obtain information and data and provide nonjudgmental feedback to teachers being evaluated. In this way teachers can make their own judgments about how to improve their own teaching.

Formative evaluation assumes a professional attitude toward individual teachers, allowing each the choice to improve his or her behavior. Formative evaluation is predicated on an underlying, three-part, philosophical belief that (a) professional teachers constantly strive for continued individual excellence; (b) given sufficient information, professional teachers can and will evaluate themselves and modify their performance as well as or better than others; and (c) the evaluation procedures provide feedback designed to assist teachers in making judgments about how they can best improve their teaching.

It follows then that the concept of self-assessment places the individual as his or her own formative evaluator. Self-assessment also carries with it the burden of the professional — that of a continuing requirement to improve the quality of one's own professional practice. If a professional formative self-assessment program is promoted or even allowed, the need for summative evaluation diminishes or disappears.

Teacher-evaluation systems are not inherently formative or summative. How the data are used determines if an evaluation system is summative or formative. If the data generated by the system are given back to teachers for their use in improvement of their teaching techniques or styles, and *never* used to make judgments about them by one who can alter placement, salary, status, tenure, or working conditions, then the system is formative. If the data generated by the system are used in any way to make judgments about placement, status, salary, or conditions, then the system is summative. Some believe that if people who can determine status, salary, conditions, even *have access* to the data, then the teacher-evaluation system should be considered summative.

A rule of thumb that the reader might consider is that "any data that can be used summatively will be." A corollary is "any and all data that were gathered as formative data and then later used summatively becomes summative retroactively in the mind of the evaluatee."

What Is Self-Assessment?

Self-assessment is not a single concept. It is composed of many elements often mixed in various ways to produce judgments about one's own teaching for the purpose of self-improvement. Self-assessment is a powerful mechanism for personal development, because the teacher who engages in it is usually highly motivated toward self-improvement. The greatest value of self-assessment is the increased instructional improvement that results from a greater insight into one's own strengths and weaknesses.

Bailey (1981) examined the origins of self-assessment and defined the concept as not merely a method, but a comprehensive approach that includes a philosophical attitude and strategies for total approach to instructional improvement. He identified seven steps for effective self-assessment: (a) gaining a philosophical overview by examining the myths surrounding teacher self-assessment; (b) the absolute necessity of using audio- and videotape; (c) identifying basic teaching behaviors or skills to be analyzed; (d) identifying verbal cues; (e) identifying nonverbal cues; (f) learning how to plan and evaluate the instructional behaviors identified; and (g) using observation forms. The entire evaluation system is designed to be executed completely and solely by the evaluatee. Bailey's book also includes a timely classified bibliography.

Three steps are necessary for effective self-assessment. First, current teaching behavior must be accurately identified. Teachers must be aware of what they are doing, both verbally and nonverbally, in the classroom. Many teachers perceive their performances as very different from reality. Thus, for improvement to occur, it is critical in the self-assessment that they become aware of their actual teaching behavior. Second, problem areas need to be identified and improved, strengths acknowledged and maintained, and new behaviors practiced. Third, new behaviors should be subjected to further evaluation to determine their effectiveness, and the whole process begins anew. Thus, self-assessment is approached as a continuous and on-going process rather than a single, one-time event.

Numerous self-assessment techniques deal with one or more of the steps listed above (data collection, behavior modification, and reevaluation). While nearly any type of evaluation technique can supply data to an instructor for the purpose of self-assessment (e.g., student ratings, colleague ratings, measures of student achievement), several methods are specifically designed for self-assessment. Each technique may be used alone, but, in practice, an actual self-assessment often uses a combination of methods, and often adds peer review components. For example, Carroll (1981) listed five major methods: self-rating forms, self-reports, self-study materials, observation of colleagues' teaching, and videotape/audiotape feedback of one's own teaching.

What Is Peer Review?

Peer review is a major component in many self-assessment programs. In these hybrid systems (evaluation systems that combine self-assessment and peer review), peers provide objective feedback to teachers to assist them in identifying teaching behaviors they may have missed and to get them to focus on new behaviors they may wish to include. Note, the use of peer review in self-assessment strategies that are formative is never done to make judgments about teachers, only to help keep them objective about their self-assessment appraisal. Peer-mediated self-appraisal (Barber 1985) is a hybrid system that combines both self-assessment and peer-review techniques. The system uses different strategies for new and beginning teachers, tenured teachers who need assistance, and tenured teachers who experience no teaching difficulties. A standard formative evaluation is applied to this latter group and deals solely with providing them with assistance in improving already good-to-excellent teaching through a combination of self-assessment and peer review techniques.

The Techniques of Self-Assessment

These techniques were culled from 269 citations and abstracts during a recent and extensive review of the literature. (See Barber, 1987, for a full listing.)

Videotape or Audiotape Feedback

This technique involves the recording of an actual, live teaching episode on either video- and audiotape or audiotape alone for the purpose of self-assessment. By far the most popular and potentially most powerful means of self-assessment, such records allow the evaluatee to see himself or herself as others do, and thus the technique is often referred to as *self-confrontation* or *self-viewing*. The method may involve microteaching, the planning and teaching of only a very brief class segment (15-20 minutes), usually to a very small group of students or peers, or the taping of a real-life classroom situation. Ideally a split screen is used so that both teacher and classroom response can be monitored simultaneously.

The technique often involves the use of some type of directed-observation instrument (e.g., self-rating forms or self-reports) or other techniques which aid the evaluatee in seeing himself or herself more objectively, eliminating some anxiety, and clarifying what it is he or she seeks to observe. Many writers claim that this very powerful self-assessment technique is most useful when evaluatees receive bona fide feedback — when they know what behaviors are expected, can see where they deviate from the norm, and when they have adopted these desired behaviors (Lewis & Barber, 1986, p. 119).

Borg, Kallenbach, Morris, and Friebe, in a 1968 report, described a mini-course model for use particularly in a university preservice program that has as its goal the improvement of 12 specific teaching behaviors related to classroom discussion techniques. The course is completed in four instructional sequences, each addressing 3 of the behaviors. Each sequence involves filmed modeling of the behaviors, a videotaped microteaching session using those behaviors performed by the evaluatee, self-viewing with observation forms, critique and replanning of the lesson, retaping of the improved lesson, and a final self-critique. The experiment was carefully monitored through the use of control groups to assure that behavioral changes were attributable to the training, and the authors supplied several statistical charts to document their results. Video feedback was found to be highly successful in affecting long-term positive change.

In a section on videotape/audiotape feedback, Carroll (1981) described a self-confrontation model and discuss three prominent techniques for evaluation using video for self-assessment: (a) microteaching; (b) interpersonal process recall; and (c) interaction analysis. Also valuable are his references to studies utilizing each of these techniques.

Self-Rating Forms

A self-rating form is a written instrument that requires the teacher to rank or grade himself or herself on a variety of teaching skills listed on the form. Often such a form is adapted from a student evaluation form. The scale may be numeric (quantitative) or verbal (qualitative). Research shows that such forms are best used in conjunction with other types of self-assessment. Two primary sources are available to learn these techniques.

Bailey (1981) wrote about the tremendous variety of available evaluation forms. Basically, such forms are simply checklists of specific behaviors, frequency of behaviors, and sequence of behaviors. As such, they are essentially objective measures. Bailey's article describes the two basic types of observation forms (those prepared by experts or professionals in the field of observation instruments, and those "custom made" by the teacher who is involved in the self-assessment process) and discusses the pros and cons of each. A section of his bibliography is devoted to materials dealing with observation instruments.

Carroll (1981) commented that there were few studies that deal with self-rating instruments. The available research suggests that self-ratings frequently have little agreement with the observations of student, colleague, and administrator ratings. The greatest agreement is shown when they are used to compare one's own performance with that of colleagues, and they are most effective when used to identify for the instructor certain unexpected discrepancies with other ratings. Included in Carroll's article is a sample faculty self-rating form and references to several other writings about such forms.

Self-Reports

Self-reports are like self-rating forms in that they are a written instrument prepared by the teacher. Unlike self-rating forms, self-reports usually use an open-ended format that consists of responses to questions related to instruction. Often these take the form of activity sheets, and are sometimes known as *brag sheets* (Lewis & Barber, 1986, p. 175). Again Carroll (1981) is a primary source for this technique.

In his section on self-reports, Carroll (1981) suggests that, despite the lack of research on their use and effects as an evaluation technique, self-reports usually provide more extensive information than a rating scale. Self-reports, therefore, serve teachers as a device for periodically taking stock of their aims, accomplishments, and shortcomings, and facilitate the setting of priorities and long-range professional goals. For best results, forms should be clear, straightforward, brief, and utilized on a regular basis (e.g., every week).

Self-Study Materials

Self-study materials are programs designed so that one may individually assess and analyze his or her teaching style, as well as investigate alternative teaching techniques and materials. Frequently these materials take the form of a *programmed text*, in other words, commentary on a topic followed by modules requiring response from the person being evaluated. Thus, the teacher reads about a certain skill or aspect of teaching and then explores his or her own approaches in that area in an attempt to gain insight into his personal style (Lewis & Barber, 1986, p. 175).

There is very little in the literature that provides direction in learning this technique. One primary source of interest is Silver & Hanson (1980). They have developed a manual for self-assessment that uses Jungian psychology, specifically Jung's theory of personality types, as its point of departure. The purpose of the manual is to assist classroom teachers in diagnosing learning styles, and in assimilating the knowledge, skills, and attitudes necessary to more effectively read and respond to student needs. The manual provides discussion of a concept, and then requires the user to respond to a series of questions based on concepts that are designed to illuminate the user's personality. These activities are grouped in three clusters: assessment of one's behavior across each of the three dimensions of personality functioning identified by Jung; thinking about oneself as a learner, how preferences affect attitude and behavior in the classroom; and thinking about one's behavior as a teacher. The basic premise of this "soul searching" is that by "knowing" oneself one can be more sensitive to students and thus, implicitly, improve teaching performance.

Modeling

Modeling refers to the observation of high quality or expert teaching examples for the purpose of imitation. Teachers can compare their own techniques with those of the expert. Their observations may be rather free and unstructured or they may be quite specifically directed toward certain aspects of the model teacher. Observation may take the form of real-life in-class viewing of the respected and/or recommended colleague or the witnessing of exemplary teaching on video or film (Lewis & Barber, 1986, p. 191).

Virtually hundreds of examples of this technique are in the literature. Almost every teacher training program provides for some type of modeling and many inservice programs for teachers involve some form of modeling. Two primary sources of reference that provide some detailed examples of the technique are Borg, Kallenbach, Morris, & Friebe (1968) and Carroll (1981).

The method of self-assessment described by Borg et al. includes a number of distinct components, one of which is modeling. As part of the instruction in learning to improve specific classroom behaviors, the teacher involved in the self-assessment watches a movie in which the behaviors are described, and then views a second film that is a microlesson utilizing that behavior. The evaluatee fills out an observation checklist that focuses his or her attention on the targeted behavior as she or he views the film.

Carroll (1981) discusses the common technique of observing one's colleagues for the purpose of making comparisons with oneself. Many opportunities for such observation are readily at hand, such as team teaching, formal programs for faculty development, and faculty colloquium sessions. He suggests that such observations are most useful for evaluation when post-observation discussion is included. He mentions the program described by Sweeney and Grasha (1978) as an example of this technique. Essentially this method is peer review that is used for the purpose of self-assessment.

Observation by an Objective Outsider

Unlike typical observations in which the observer actually performs the evaluation, this method is done by a disinterested outsider who simply records his observations (e.g., types of activity in the classroom, their frequency, and their sequence). The observer then submits the data to the evaluatee for his or her own self-analysis of the results. This method is sometimes called *systematic observation* (Lewis & Barber, 1986, p. 191).

Again, this technique is described in volume in the literature. There are large numbers of observation schedules, checklists, and training programs that teach one how to be an objective observer. Most of these references exist as ways of gathering data to perform a summative evaluation either more completely or objectively. If one, however, turns the data over to the evaluatees to be analyzed

for increasing the quality of their own instruction, the process becomes formative and a powerful self-assessment tool.

Of the many examples available we consider Medley, Coker, and Soar (1984) as a primary source if it is used to provide data back to the evaluatees for their own use in self-assessment. In this book the authors provide a strategy that demands that the objective observer remain objective by attending to a coding system while observations are made and having the scoring conducted after all observations are concluded.

Questionnaires

Questionnaires are typically given to current and former students. In formative self-assessment it is the person being evaluated who analyzes the data obtained. The questionnaires are merely regarded as another source of information (Lewis & Barber, 1986, p. 205).

Interviews

The evaluatee may conduct interviews with appropriate people as another way of obtaining data for a self-assessment. Most typically, interviews are conducted with current and former students (Lewis & Barber, 1986, p. 205).

Cranton (1978) provides a case study that shows how interviews can be used with past and present students and outside others. Interviews can be open-ended or deal only with previously determined topical areas.

Use of a Consultant, Expert, or Peer

In this method a teacher who is in the process of self-assessment enlists the assistance of a consultant or expert. The consultant does not perform the evaluation, but merely aids the evaluatee in setting up and carrying out the evaluation. Complete confidentiality is maintained. A combination of strategies that involves the use of a consultant is found in Cranton's (1978) study.

Another combination approach is found in Dugas (1967). This approach describes a workshop for improving classroom techniques in a high school French class. The basic evaluation tool used was videotaped microteaching sessions. However, an expert consultant was an integral part of instructing the teacher in how to evaluate his or her videotaped performance.

Comparison to Standards

Comparison to previously established standards is implicit in virtually every form of evaluation. However, a few writers mention specifically the comparison of the performance of teachers in executing their responsibilities to such things as their job descriptions, written goals that they have established for themselves,

and so forth (Lewis & Barber, 1986, p. 215). One excellent example of this strategy is found in a Gwinnett County (GA) evaluation system (1985). Although quite general in its approach, and designed mainly for summative purposes, evaluation is explicitly achieved by comparing a teacher's performance to: (a) the "Goals Plan" (i.e., goals identified to meet the needs of the school system, the school, or the evaluatee personally) and (b) the person's job description.

Hybrid Systems

In actual practice, examples of a single technique of self-assessment used alone are seldom found. Rather, one finds mixtures of self-assessment and peer review strategies, hybrids if you will.

The basic premise behind hybrid techniques is that no single type of evaluation can adequately meet the needs of all people involved in any evaluation process. Each type of evaluation suffers inherent problems. By combining approaches both issues can be addressed and resolved. Weaknesses of certain techniques can be ameliorated or balanced by combining them with other techniques. The purpose is to allow the evaluatee a sense of comfort and trust along with improvement in teaching while at the same time meeting the demands of the administrator and public for objectivity.

Few hybrid systems were found in the literature search. Of those found, most followed a similar pattern of steps to produce a hybrid system of self-assessment and peer review.

Step 1: Identification of Purpose, Goals, Objectives or Intent

This step is done by the teacher either alone or with the assistance of the peer or other objective, nonthreatening, reviewer.

Step 2: Preparation of a Teaching Plan

This step is almost always done by the teacher alone. Here the teacher identifies the where, what, why, when, and techniques and strategies of the lesson to be taught.

Step 3: Preobservation Conference

This step is conducted by the teacher and the reviewer. Its purpose is for the teacher to clearly identify the behavior the reviewer is to observe, how records are to be kept, how feedback is to be given, and what conditions of confidentiality are to be met.

Step 4: Teach the Lesson, Record the Teaching, Obtain the Material Agreed to in the Pre-conference

In almost all cases this step uses video recordings of teaching so the teacher can observe his or her teaching and perform his or her own self-assessment. It also allows the reviewer to observe the teaching at his or her leisure.

Step 5: Teacher Self-Assessment of Teaching Behavior

In this step the teacher uses one or more self-assessment techniques (usually video, self-rating forms, and student interviews) to determine the quality of teaching and also to determine if the goals in step one were satisfied. The teacher then writes out his or her findings and conclusions.

Step 6: Provision of Data to Reviewer

In this step the reviewer is provided with all the data agreed to in the pre-observation conference, step three, and is provided with the teacher's findings and conclusions drawn in step five by the teacher. The peer reviewer's task is to determine the degree of correctness of the teacher's findings and conclusions from the data provided. Generally the reviewer is *not* to determine the quality of teaching. The reviewer's findings and conclusions are written and provided to the teacher.

Step 7: Post-Observation Conference

This step involves a face-to-face (or telephone) conference between the teacher and the reviewer to discuss any discrepancies that may exist between the teacher's and reviewer's findings and conclusions.

Step 8: Resolution of Discrepancies

This step is conducted by the teacher to determine if modifications in teaching techniques, strategies, procedures, and materials may be required after analysis of the discrepancies found between the teacher's findings and conclusions and those of the reviewer. This step might also be considered to be the beginnings of another round of steps one through eight.

Hybrid techniques seem to be a relatively new approach to teacher evaluation, yet seem to be gaining in popularity. The literature contains little information regarding this method. Two references clarify the technique.

Peer mediated self-appraisal (PMSA) (Barber, 1985) is an evaluation system that combines both peer review and self-appraisal, as successfully applied in public school settings, in an attempt to satisfy board goals, district goals, and individual building goals. The method should lead to improved quality of instruction with the insurance of a clear objective check using peer review, while attempting to protect the teacher from biased evaluators, should that occur.

The system accounts for new and beginning teachers, contracted teachers who are having difficulty, and contracted teachers who are not experiencing problems. The standard formative evaluation applies to this latter group and deals solely with the processes and strategies of teaching. It is accomplished through a combination of peer- and self-review processes: (a) tenured teachers undergo self-assessment annually and peer review periodically; (b) they are responsible for maintaining their own records, which are submitted during the standard professional (summative) review, only to show that they were attempting to improve their own teaching; (c) the review is conducted by the teacher's supervisor and occurs at least once every three years; (d) the process involves gauging performance against previously identified obligations.

In summary, the system attempts to use formative evaluation through self-assessment and peer review to increase the quality of teaching and also (once every three years) use summative evaluation to satisfy the demands of the public for accountability. Thus, a teacher who fails to meet the summative evaluation criteria is placed on "intensive assistance," during which he or she receives inservice training and counseling in an effort to meet competency standards. The grand plan of PMSA allows that during any year two-thirds of the faculty will be involved in formative evaluation on their own terms and the other one-third in a summative standard professional review.

Forbes & Burkhardt (1988) describe how a school district in central Indiana developed a way for teachers and administrators to develop professionally as they helped students improve academic achievement and adapt to a new desegregation plan. Emphasis is on the TESA (Teacher Expectations and Student Achievement) program, which serves as a positive introduction to, and a prerequisite for, other staff development programs and activities. Experiences with the TESA program empowered teachers to change their classroom management style, to participate in peer observation and evaluation, and to demand additional training to further enhance their skills. Please note, the author has not been able to identify a single inservice workshop designed to teach teachers the techniques of self-assessment and peer review as of the writing of this chapter. It is important then to note that Forbes, recognizing the absence of such needed training programs, has developed and is currently field testing a workshop that teaches the hybrid system steps noted earlier.

Problems with Self-Assessment

Up to this point, the author has identified formative evaluation, self-assessment, and peer review techniques in a very positive light. They are, in fact, extremely important techniques that have the potential to improve teaching behavior more efficiently and quickly than any other techniques known at this point. They are, however, not perfect — there are limitations to these techniques. In an effort to balance this work, some of those limitations are listed below.

1. Lack of objectivity.
2. Lack of accuracy and reliability; not a meaningful measure of competence or performance.
3. Individuals tend to regard themselves as proficient, and honest/objective evaluation is difficult — some consider it impossible.
4. Evaluation may become a form of self-justification (e.g., an incompetent person may not realize he or she is performing at an unsatisfactory level).
5. Mediocre teachers tend to be less accurate in self-assessment than superior ones.

6. A tendency to depend on the consultant, if one is used, to carry out too much of the evaluation.
7. Inherent potential for self-incrimination if the results are used summatively.
8. Difficulty in quantifying the assessment.
9. A tendency to focus on cosmetic things (such as hair, mannerisms, etc.) when using video, rather than on matters of substance.

Conclusion

Self-assessment as a part of teacher evaluation can be the most powerful force available to us as educators in our never-ending quest to improve teaching behavior, strategy, and technique. Self-assessment, where it is allowed to flourish in a nonpunitive environment, supports an internal drive in all professional teachers to constantly improve their teaching behavior. This drive, once awakened, demands more improvement and better performance and is far more powerful than external requirements or demands placed upon us by others.

Peer review, nonthreatening collegiality, modeling, and peer sharing of quality teaching are extremely effective ways of improving teacher performance. Given their choices, professional teachers constantly request these opportunities because they see them as more effective ways of improving their own teaching. Combining self-assessment and peer review into hybrid systems has the potential of improving teacher performance so effectively that, if used wisely, could greatly reduce the need for punitive summative forms of teacher evaluation.

References

- Bailey, G. D. (1981). *Teacher self-assessment: a means for improving classroom instruction*. Washington, DC: National Education Association.
- Barber, L. W. (1985). *Improving teacher performance: Formative evaluation*. Bloomington, IN: Phi Delta Kappa.
- Barber, L. W. (1987, January). *Teacher evaluation*. A report submitted to the Teacher Quality Program, Indiana Department of Education.
- Borg, W. R., Kallenbach, M. M., Morris, M., & Friebe, A. (1968). *The effects of videotape feedback and microteaching in a teacher training mode*. Berkeley, CA: Far West Laboratory for Educational Research and Development.
- Carroll, J. G. (1981). Faculty self-assessment. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 180-200). Beverly Hills, CA: Sage Publications.
- Cranton, P. A. (1978). *Evaluation of teaching for improvement* (Report No. 1). Montreal: McGill University. Presented at the University of British Columbia, March 6, 1978.
- Dugas, D. G. (1967). Micro-teaching: A promising medium for teacher training. *The Modern Language Journal*, (March), 161-166.
- Forbes, R. H. & Burkhardt, G. E. (1988). *Decatur township, Indiana: A staff development model (featuring the TEESA program)*. Bloomington, IN: Phi Delta Kappa.
- Gwinnett County Public Schools. (1985). *School-based administrators performance evaluation*. Lawrenceville, GA.
- Medley, D., Coker, H., & Soar, R. (1984). *Measurement-based teacher performance evaluation*. Lawrenceville, GA.

- Lewis, S., & Barber, L. W. (Eds.). (1986). *Improving your own instruction: self-assessment and peer review*. Bloomington, IN: Phi Delta Kappa.
- Silver, H. F. & Hanson, J. R. (1980). *Teacher self-assessment*. Moorestown, NJ: Hanson, Silver & Associates.
- Sweeney, J. M., & Grasha, A. F. (1978). Improving teaching through faculty development triads. *Improving College and University Teaching Yearbook*. Corvallis, OR: Oregon State University Press.

14. Using Student Test Scores to Evaluate Teachers

Gene V Glass

Introduction

In 1981, Jason Millman presented a chapter entitled "Student Achievement as a Measure of Teacher Competence" in his first edition of the *Handbook of Teacher Evaluation*. His approach was analytic: he broke down the global problem into its constituent systems; he defined terms and illustrated concepts and generally equipped the reader with the understanding of the elements of an evaluation process that would seek evidence of teachers' accomplishments in the learning of their students. This author could not improve on Millman's analysis and will not attempt to do so; the reader who desires the analytic account of the problem cannot do better than read Millman's chapter in the first edition.

The intent of this chapter is different. Here, the author wishes to present a synthetic account of how the elements Millman identified are likely to fit together into an actual system for evaluating teachers. The whole is not merely the sum of its parts, as the shop-worn tag goes; but more importantly, the machine that is meant to do the world's work must be observed in the actual world where designs meet reality. When technical desiderata meet practical and political limitations, a truth of its own special sort is created. The author has attempted to describe how a system of teacher evaluation that makes genuine use of student achievement data will work when attempted in contemporary schools. The issue of "pupil performance in teacher evaluation" needs to be discussed in a way that has not been separated from the many other pressures that shape a personnel system of assessment and rewards. The issue cannot be judged apart from how it exists in real places. The validity of such a process of evaluating teachers, its economy, its power to engender belief and win acceptance, depend on how it fits with many other realities of the teaching profession. What now must be addressed is how this notion gets applied in a context as complicated as

AUTHOR'S NOTE: These persons served as consultants on the preparation of this chapter: Mary Catherine Ellwein, University of Virginia; Mark E. Fetler, California State Department of Education; Del Schalock, Oregon State System of Higher Education.

education. How is the idea transformed as it moves from the statistician's mind to the real world? How does it fare when the tradeoffs and balances are struck? Is the concept of evaluating teachers by student progress trusted? Is it more a symbol of certain values that the community requires be honored than a reality of school personnel practices? Presenting hypothetical possibilities will not answer these questions, rather they are answered by portraying complex realities. The description to be presented is based on the actual experiences of a half-dozen school districts in which student achievement scores of tests were clearly tied to the evaluation of teachers for pay raises. Persons who viewed the process from many different sides were interviewed for the purpose of assembling this narrative; in addition, documents and publications about the programs were read and analyzed. Facts from the different sites merged naturally into a single coherent picture of what probably happens when teachers are evaluated and rewarded on the basis of student achievement measures. Consequently, the author has presented a single composite of the half dozen sites. The name for the district is fictitious.

The Composite Sketch

The Montview Unified School District initiated a teacher merit pay program in 1975. The administrative staff had reviewed the merit-pay programs of many school districts around the country. They read dozens of plans that referred vaguely to teachers being evaluated by pupil achievement; but never could they see where this promise was kept in a direct and systematic way for a entire district. They came to regard such rhetoric as mere lip-service paid to pupil growth, and they vowed to do better. The merit-pay program they designed provided financial bonuses in amounts up to \$2,500 for teachers who achieved the goals identified by the superintendent and his staff who designed the program, without advice or consent of the local NEA and AFT affiliates. Even this simple principle of incentive and reward did not go unchallenged by community leaders. Critics argued that teachers were adequately compensated and that a system of bonuses could cause teachers to elevate personal gain above pupil well-being; proponents of the merit-pay idea argued that pupil well-being could be embodied in the criteria for merit bonuses, so that no conflict of values need arise. Moderates argued that teachers should receive bonuses in the form of fees for professional workshops instead of taking them as income. The Montview school administration backed this proposal briefly before its quick death at the hands of an aroused teachers' organization.

In the end, an ambitious superintendent's desire to make a reputation combined with a growing sense that American education was facing a crisis of international proportions overcame any doubts that might have remained.

Setting the Standard for Merit Reward. Bonuses for teachers who achieved pupil growth goals were a big part of the Montview Teacher Merit Pay (TMP)

program. Though it was only one of several criteria for earning extra money, assuring pupil mastery of basic skills was politically crucial for winning public approval of the TMP program. In the spring of 1974, the superintendent revealed the details of the TMP plan:

Teachers will receive bonuses for (a) improving their own attendance in the classroom; (b) meeting goals for student academic growth; (c) teaching a subject in which there is a critical shortage of teachers (math, science, and special education); (d) teaching in a school with a high percentage of minority pupils. Teachers would have to apply to be reviewed for TMP bonuses, with applications due by June 15 of the summer following the year under review. In this way, teachers who might not be in line for a merit bonus could avoid the embarrassment of being publicly turned down; if asked by a colleague, they need only say that they forgot or did not have the time to apply.

A 50-page manual of procedures was released in June that set out the details of the TMP bonus program. Teachers serving in schools with more than 50% minority enrollment would receive an additional \$500 pay for the 1975-76 school year. Teachers of math, science (at junior high and high school levels) and special education would receive an additional \$500 bonus. The teacher attendance criterion was the focus of particular attention. Normally teachers were eligible for ten sick-leave days per school year. Substitute teachers cost the district about \$50 a day. Teachers who missed fewer than five days during the school year would qualify for a financial bonus; for each day under five days total absences, the teacher was to receive \$100. Hence, a teacher with no absences during the year would receive \$500, one absence was worth \$400, two absences earned \$300, etc.

The Pupil Achievement Growth Standard. The regulations governing the pupil growth bonus were the most complex. At its first meeting, the administrative staff proposed a quota system for allocating bonuses to teachers for outstanding pupil progress. By whatever means ultimately chosen, teachers would be ranked from highest to lowest class average achievement gain during the school year, and the top half of the teachers would receive a \$500 bonus. Nothing else would truly motivate teachers to excel, it was reasoned. Protests to this plan from teachers were immediate and loud. They promised to reject any plan that even smacked of a quota system. The recent experiences of a neighboring state were invoked to support the teachers' case; when the Legislature of Tennessee enacted a career-ladder program for teachers, they stipulated in the act that rewards must not be allocated on a quota basis; all who qualify must be rewarded, and teachers must compete against an external standard instead of being pitted against each other. The Montview schools administrative staff had lost round one. A week later they presented their second attempt at a student progress bonus plan.

First, it was recognized that academic progress in some areas of the school curriculum was impossible to measure in any practical way. A number could be put on reading achievement for a class — though many would dispute its meaning — but measuring progress in art, physical education, and the like was an entirely different matter. Thus, an important element of the TMP plan from the beginning was the decision to base the \$500 bonus not on the progress of an individual teacher's class but instead on the progress of the pupils in the entire school. Moreover, it was clear to the teachers' organization that academic progress in basic skills at the secondary school level was less relevant than at the elementary level. Before school began in the fall, the regulations were revised: "All teachers in a school in which 75% of the teachers meet their pupil-growth goal will receive a \$500 bonus for the year. The pupil-growth goal will be defined in terms of standardized test achievement in reading and math at the elementary school level, and in terms of criterion referenced test mastery at the secondary school level." The superintendent's staff felt that the negotiated regulation was the teachers' way of reducing competition, though they were willing to discuss it with the teachers as if it were a technical problem of quantifying difficult-to-measure goals.

The superintendent and his staff saw clearly that the teacher attendance criterion could potentially save the district tens of thousands of dollars in substitute teacher fees. Indeed the criterion was designed to do just that, though publicly one heard far more about the need to provide continuity in the instructional program than of the cost savings by reducing substitute teacher use. So at the last stage in drafting the regulations for the TMP program, they reinforced the importance of the attendance criterion and made the pupil-growth criterion dependent on it: only those teachers who met the attendance criterion of five or fewer days absent would receive the pupil growth bonus. The TMP bonus system could reward a teacher with as much as \$2,000 additional pay in a school year. For example, a teacher of math with one day absent from a ghetto school assignment where the school met its student-growth goal would receive the maximum \$2,000 bonus. An elementary school teacher in a nonminority school who missed seven days work would receive no bonus, regardless of the academic progress of the school or the teacher's class.

Refining the Pupil-Growth Standard. The administrative staff set to work in the fall to define the student-growth criteria that would qualify schools for teacher bonuses. They recognized immediately that teachers work in dissimilar settings. Some account must be taken of pupil ability, home circumstances, past school experiences, and the like in setting a pupil growth criterion. Any criterion such as "The pupils in your class will average above the 50th percentile on the Metropolitan" would be seen as unfair to teachers of large classes of slow-track pupils or of pupils in transient neighborhoods where half the students present at the end of the year were there when the school year began. The staff flirted temporarily with the notion that pupils in each grade could be randomly assigned to the teachers, thus assuring that each class started the year somewhere near

equality on achievement, learning potential and all the rest. This idea was rejected without serious consideration when it was realized that the differences between schools were surely as large as any differences within them and no one could imagine randomizing across schools, and when the staff remembered that parents' preferences for particular teachers were very strong.

The first serious suggestion for a criterion was made for the elementary grades: the school must show a gain on standardized tests of one grade-equivalent year to qualify for the teacher bonus. The district research and evaluation office did a quick check of how many elementary schools met this criterion in the two previous years, and discovered that 60% of the schools showed gains of a year or greater. When the teachers' organization learned of these data, they protested that the criterion was unrealistically severe. The administration backed off. A consultant from the local university advised the district to adopt a method of linear regression residuals to measure teachers' contributions to pupils' academic growth. The consultant showed how entry level achievement, ability and socioeconomic level could be taken into account and used to adjust each pupil's year-end achievement score so it would reflect growth from an equal starting point. The research and evaluation department's resident statistician protested that the pretest measures of ability and achievement were less than completely reliable, that motivation wasn't captured accurately in any of these correction variables, and that when all the adjusting was done, the administration would still face the problem of deciding how large an adjusted gain must be to qualify for a bonus. The teachers' organization rejected the consultant's advice as too esoteric to explain to the public or to teachers. Finally, several persons suggested almost simultaneously that if the school building's standardized achievement-test grade-equivalent gain was calculated for the previous year and if two months were added, a criterion that was clear and fair would result. The objection that the addition should be two months in richer schools and one month in poorer schools was quickly overridden in the interest of arriving at a solution.

At the secondary school level, the derivation of a pupil-progress criterion was less convenient. Few of the teachers in a school above grade seven are directly involved in teaching basic academic skills. Standardized achievement tests would not solve the problem. The administration recommended that each teacher develop an end-of-year criterion referenced test and set a criterion of success in terms of the percent of items answered correctly. If 80% of the teachers in the secondary school met their criterion, all the teachers in the school would qualify for the merit bonus.

The First Year: Everybody Wins. The criteria having been set, the TMP program ran for its first year without any serious complications. By June 15 of 1976, the data were analyzed and the bonuses paid. One-third of the teachers received \$1,000 in bonuses for teaching in subjects of critical need in minority schools. Of the 600 elementary and secondary teachers in the district, 85% reached the pupil-growth criterion. Only three elementary schools did not meet the "Last year's gain + 2 months" criterion for pupil growth. The teachers in one

of the schools protested that the pupils had learned about the bonus feature and purposely bombed the posttest to make the teachers look bad. A second complaint was registered: some classes had too many bright pupils and enjoyed an unfair advantage. The research and evaluation office agreed to throw out the bottom 5% and top 5% of scores in any school and recalculate the gain. The school then achieved the criterion and so did one of the other elementary schools. This "throw-out-10%" rule became a standard part of the measuring procedure in later years to guard against sabotage and improve the appearances of comparability. Against the recalculated standard, 97% of the teachers qualified for the student-growth bonus; only 3% of the teachers worked in a school that didn't meet the criterion. Virtually all teachers qualified solely on the basis of student progress for a \$500 bonus, but about 70 of them lost it because they had been absent from work more than five days during year. As a result, about 85% of the elementary and secondary teachers received \$1,000 bonuses for attendance and pupil growth.

Only one teacher in the district refused to volunteer for evaluation under the TMP program. Regina Tyle taught English Composition at the junior and senior levels. Tyle was active in the Montview Teachers' Association, and her reputation as an excellent teacher was well known. Assistant Superintendent for Instruction Ralph Marshall called her in when it became apparent that Tyle was not going along with the program.

Marshall: "Why aren't you in it, Regina; you're one of the best? It's a sure \$1,000."

Tyle: "I don't believe in it. Where's a test for creativity? Where's the test that gets at what I'm trying to give these kids?"

Marshall: "Look, just make up something yourself. It doesn't have to be any big deal; we just want something we can point to and say, 'Yes, she met the criterion.'"

Tyle refused, on principle. Within 18 months, Tyle assumed the presidency of the state teachers' association and left the district.

The first year of operation of the TMP program was regarded as a success. Standardized test data for the Montview district were up more than two months above the level of previous years for the elementary grades; at the secondary grades, no difference was apparent in test performance, but administrators reported a new seriousness of purpose among teachers and more businesslike classrooms. One set of figures permitted no disputed interpretations. The average number of days teachers missed work in 1974-75, the year before the TMP program, was 12.5; during the initial year of the TMP program, the average teacher missed only 3 days. In 1974-75, the Montview Unified School District paid out \$385,000 for substitute teacher fees for absent teachers. In 1975-76 under the TMP bonus plan, the average teacher was absent 3 days during the school year, which amounted to \$100,000 in attendance bonus money (since

15% of the teachers missed more than 5 days and failed to qualify for the bonus) to the teachers and \$90,000 in substitute teacher costs. Indisputably, the TMP attendance bonus plan saved the district \$185,000 in personnel costs. The administration hailed the TMP program as a tremendous success. Superintendent Stevens's stock rose; his phone began to ring with inquiries from around the country.

The Second Year: New Efforts to Reach the Standard. The second year of the TMP program began amid growing realization by the administration and the school board that the pupil-growth feature was costing the district \$255,000 per year in bonus money and yet it was nearly perfectly correlated with the attendance bonus. Virtually all teachers qualified for the pupil-growth bonus; it was denied only to those who missed too many work days. The \$255,000 for the pupil-growth/attendance bonus combined with the \$190,000 in direct attendance bonus money put the cost of the TMP bonuses at \$445,000, which was \$70,000 more than the district spent for substitute teachers the year before the TMP program began. The TMP bonus system started to come under heavy fire at board meetings in the early fall of 1976-77. Having announced the bonus program with great publicity and having declared it a success, the administration and the board could not dissolve it without embarrassment, and besides, the test scores were up two months above past levels, at least at the elementary level. Nonetheless, the deliberations were tense as the board undertook its major item of business on the fall agenda, the revamping of the teacher salary schedule. Three month's deliberation and argument produced a new salary schedule that reflected a slightly reduced rate of growth in teacher salaries for the Montview district compared with the recent past rate and neighboring districts, which had begun to complain to Stevens and the Board that the historically higher Montview schedule was placing pressure on them to raise their schedules. The clinching argument for decelerating the growth rate in the salary schedule was the existence of the TMP bonus program, which was adding about \$1,000 to the average teacher's salary each year. In effect then, by the middle of the second year of the TMP merit program, the Montview Unified School District had merely designated a portion of teachers' salaries as attributable to meritorious attendance on the job and a second portion as attributable to meritorious enhancement of pupil academic progress. It would have been political impossible to sell the public on the notion that teachers should receive a \$1,000 bonus for merely showing up to do a job for which they were already being paid; so \$500 was for attendance and \$500 was for improved pupil progress. Thus, pupil achievement came to serve as a proxy for improved teacher attendance.

By the middle of the second year, teachers began to see the difficulties of continuing to meet the pupil-growth criterion. Teachers at the secondary level seemed little concerned. At meetings between secondary teachers and curriculum specialists, the word was given out that last year's criterion-referenced mastery goals would have to improve by at least 2% for teachers to qualify for the bonus. The teachers insisted that they be given the opportunity to make minor

revisions in their tests to improve reliability and validity. This eminently reasonable request was granted and nothing more was heard. The data reported by secondary teachers in June showed that all secondary schools had once again reached the pupil-academic-progress goal.

At the elementary level, things were different. By the rules specified at the start of the TMP program, an elementary school that made 8 months grade-equivalent gain in 1975-76 would have to make 10 months gain in 1976-77 to earn the bonus. The "throw out the top and bottom 10%" procedure had already been applied to the 1975-76 data so that would not help improve the picture for 1976-77. An inquiry from the teachers to the administration about switching from the Metropolitan to the Iowa Test was met with promise of rejection were such a request to be made. Several teachers remarked that their 1975-76 scores had been unfairly depressed by children transferring into their class after Christmas. They requested that for 1976-77, no scores be included for pupils who spent less than one full semester in their school; the administration granted the request. Uncertain that this new proviso alone would result in a significant increase in the 1976-77 gain as compared with the 1975-76 gain, the teachers set to work by January to ensure that their pupils would be ready for the Metropolitan on its April visit. Over the next three and one-half months, testlike worksheets in language arts and math made increasingly frequent appearances in the Montview elementary school classes. At grades two and three, a series of units on addition and subtraction were hurriedly rewritten to change vertical display of math facts to the horizontal, since the Metropolitan employed the latter format (perhaps as a means of reducing paper costs). By April, the pupils were ready. As in the first year of the TMP program, teachers administered the standardized tests in their own classes. Many were helped by student teachers from the local university. Student teachers asked their professors at the university what they thought of the practice of "talking the kids through the Metro." Everyone in the class understood what this meant. When asked by the professor whether they would do such a thing if their salary depended on it, no one demurred.

The June data for the elementary schools were encouraging. The average gain for the district was 11.3 months grade-equivalent units. Only one elementary school failed to make the 10 months growth criterion, but since it was a high minority enrollment school, all the teachers received a \$500 bonus for service in a minority school. In the second year of the TCL program, no teacher failed to receive at least \$500 bonus money, and the average was \$1,100; 90% of the teachers received \$1,000 for the attendance and the pupil-growth bonuses.

Years Three and Beyond. Year three of the TMP program ran without any significant changes. The elementary-pupil-growth criterion had risen to 6 months over where it began in each school, but all elementary schools reached it. In year four, the standard was 8 months above its original level, and the teachers began to complain that the criterion was becoming unrealistic. The school board heard the teachers' organization's appeal for relief from the increasingly severe pupil-progress criterion, but the steady outlay of over a half million dollars per

year for TMP bonuses and growing pressure on the board to bring the basic district salary schedule more in line with those of neighboring districts forced them to hold the line on the "Last year's growth + 2 months" criterion. In year five, the elementary schools were beginning to bump against the ceiling of the Metropolitan achievement test. Eight of the 20 elementary schools failed to make the pupil-growth standard. The average TMP bonus in the district dropped to \$700. The teachers began to complain vigorously.

Before the start of the 1980-81 school year, the teachers' organization secured an agreement with the administration that teachers could choose which of the various levels (Primary, Intermediate I, Intermediate II, etc.) of the Metropolitan would be given in their class. Most of the teachers elected to have their pupils tested on the level of the test one step above that which they had administered for the last five years. The typical criterion in year six was 14 months grade-equivalent units. In June of 1981, all but one elementary school once again made the pupil-growth criterion. But the success was short-lived. The criterion of 16 months in 1982 was reached by only two-thirds of the schools, and in 1983, a third of the schools exceeded the growth standard of 15 grade-equivalent months. Academic excellence had topped out in the Montview Unified School District, and elementary teachers's bonuses, now obviously paid for low absences, dropped to an average of \$600.

Administrative staff and principals watched as the amount spent on bonuses in the TMP program shrunk from about \$600,000 in 1976 to about \$350,000 in 1983. Teacher attendance had reached the level of about three absences per year in the first year of the program and remained at that level. But the number of teachers rewarded for outstanding pupil growth dwindled each year under the increasing pressure of producing greater growth rates year-in-year-out. During the summer of 1983, the administrative staff in collaboration with the building principals designed a School Collaborative Productivity Plan that entailed joint goal setting by teachers and the building principal in such areas as teacher attendance, pupil attendance, pupil academic growth and energy conversation. The SCPP quickly superseded the TMP, as it subsumed the major elements of the TMP, but in addition now provided incentives for principals. Under the SCPP, principals could earn bonuses up to a maximum of \$9,000; teachers could earn a maximum bonus of \$1,000. At the end of year one of the SCPP, the calculations showed the average teacher receiving bonuses of \$660, and the average principal earning bonuses of \$8,300. The net cost of the SCPP program in its first year of operation was about \$600,000. Bad times lay ahead for the principal bonus for student growth. Eventually two principals were detected altering answer sheets before they were sent for computer scoring; the principals were given classroom teaching assignments the next year and permanent replacements were installed in their administrative positions. The testing company selling scoring services to Montview devised a technique for automatically detecting unusually high rates of erasures of wrong answers. The Testing and Evaluation Office felt this would solve the problem of cheating.

In the Spring of 1984, the State Education Agency analyzed and reported the results of its state-wide testing of elementary pupils on the Iowa Tests. The Montview Unified School District elementary school program was placed on probationary status for scoring significantly lower than expectation.

The vague air of scandal that hung around the TMP program had become a political liability. Teachers spoke out more openly about its drawbacks. In the spring of 1985, Superintendent Stevens left his position to assume the superintendency of a large school district out of state. His successor suspended the TMP program.

Some Reflections on Teacher Evaluation with Student Achievement Data

This description focused on systems of evaluation that related teacher performance to pupil achievement. The goal of the evaluation was to allocate rewards to teachers — pay raises. Such high-stakes evaluation subjects the system to severe stress; there is little surprise, perhaps, that it crumbles under the pressure. Some persons imagine a more modest role for teacher evaluation using student achievement. Perhaps student-progress data could be used to diagnose and correct weaknesses in teaching style or procedures. The author has not discussed this approach because he has not seen it in operation and doubts that it would work. While there is much to recommend evaluation of teaching through direct observation by supervisors and other experts, inferring errors of teaching procedure from students' test performance would be too dubious a practice to sustain any serious interest. Others, still holding on to hope for what seems a good idea, may argue that student achievement still can contribute to the evaluation of teacher effectiveness, without specifying in what context, for what purpose, or with what particular decisions in mind. Teachers should beware of such obfuscation. Student-achievement data cannot tell teachers how to teach; such data are not viewed as credible for distinguishing good teachers from bad ones; and data once gathered will tend to be used.

Few still believe that facts are simply facts. What are called facts are the result of long, complex processes of interpretation. Even more are general conclusions the obvious product of a process of interpretation that goes far beyond "the data" and taps the beliefs and world view of those who adduce them. Some may read the above account and come away convinced that it speaks only to the urgent need to make all tests criterion referenced or to step-up procedures for detecting and punishing deception. Others may see nothing but reasons to abandon what they think was a bad idea. The reflections that follow are rooted in the author's experience with similar endeavors; they are consistent with his attitude toward most attempts to make persons and organizations accountable for performance on tests (Glass & Ellwein, 1986; Ellwein, Glass, & Smith, 1988).

Using student achievement data to evaluate teachers . . .

1. . . is nearly always undertaken at the level of a school (either all or none of the teachers in a school are rewarded equally) rather than at the level of individual teachers because
 - (a) no authoritative tests exist in most areas of the secondary school curriculum, nor for most special roles played by elementary teachers;
 - (b) teachers reject the notion that they should compete with their colleagues for raises, privileges and perquisites;
2. . . is always combined with other criteria (such as absenteeism or extra work) which prove to be the real discriminators between who is rewarded and who is not;
3. . . is too susceptible to intentional distortion and manipulation to engender any confidence in the data; moreover teachers and others believe that no type of test nor any manner of statistical analysis can equate the difficulty of the teacher's task in the wide variety of circumstances in which they work;
4. . . elevates tests themselves to the level of curriculum goals, obscuring the distinction between learning and performing on tests;
5. . . is often a symbolic administrative act undertaken to reassure the lay public that student learning is valued and assiduously sought after.

References

- Brandt, R.M., & B.M. Gansneder. (1987). *Teacher incentive pay programs in Virginia*. Charlottesville, Virginia: University of Virginia.
- Darling-Hammond, L. & Wise, A. (1983). Teacher evaluation in the organizational context: a review of literature. *Review of Educational Research*, 53, 285-328.
- Ellett, C.D., & J.S. Garland. (1987). Teacher evaluation practices in our largest school districts: Are they measuring up to state-of-the-art systems? *Journal of Personnel Evaluation in Education*, 1, 69-92.
- Ellwein, M.C., & Glass, G.V. (1986). Reform by raising test standards. *Evaluation comment*, 10(1), 1-6. CRESST, UCLA, Los Angeles.
- Ellwein, M.C., Glass, G.V. & Smith, M.L. (1988). Standards of competence: Propositions on the nature of testing reforms. *Educational Researcher*, 17, 4-9.
- Florio, D.H. (1986). Student tests for teacher evaluation: a critique, *Educational Evaluation and Policy Analysis*, 8, 61-62.
- Haertel, E. (1986). The valid use of student performance measures for teacher evaluation, *Educational Evaluation and Policy Analysis*, 8, 45-60.
- Hatry, H.P. & J.M. Greiner. (1985). *Issues and case studies in teacher incentive plans*. Washington, DC: The Urban Institute Press.
- McNeil, J.D. (1981). Politics of Teacher Evaluation. In J. Millman (Ed.), *Handbook of Teacher Evaluation*, (pp. 272-291). National Council on Measurement in Education. Beverly Hills, CA: Sage Publications.
- Miller, L. & E. Say. (1982). This bold incentive pay plan pits capitalism against teacher shortages. *The American School Board Journal*, 169 (Sept.), 24-25.
- Millman, J. (1981). Student performance as a measure of teacher competence. In Millman, J. (Ed.), *Handbook of teacher evaluation*, (pp. 146-166). National Council on Measurement in Education. Beverly Hills, CA: Sage Publications.
- Moore, N. (1987). *Regression residual approach to deriving a teacher effectiveness index using student achievement: Four variations*. Tempe, AZ: College of Education, Arizona State University.

- Ryan, J.M., & G.G. Rowzie. (1987). *A manual for the student achievement component of the teacher incentive program*. Columbia, SC: College of Education, University of South Carolina.
- Say, E., & L. Miller. (1982). The second mile plan: incentive pay for Houston teachers, *Phi Delta Kappan*, 64, 270-271.
- Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of Teacher Evaluation* (pp. 244-271). National Council on Measurement in Education. Beverly Hills, CA: Sage Publications.
- Scriven, M. (1987). Validity in personnel evaluation, *Journal of Personnel Evaluation in Education*, 1, 9-23.
- Soar, R.S., D.M. Medley, & H. Coker. (1983). Teacher evaluation: A critique of currently used methods, *Phi Delta Kappan*, 65, 239-246.
- Wise, A.E., & L. Darling-Hammond. (1984). Teacher evaluation and teacher professionalism. *Educational Leadership*, 42, 28-31.
- Wise, A.E., L. Darling-Hammond, M. McLaughlin, & H.T. Bernstein. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: The RAND Corporation.

15. The Schoolteacher's Portfolio: An Essay on Possibilities

Tom Bird

Discretion might have dictated another title for this discussion of documentation procedures and their collected products. *Schoolteacher* seems to some persons to be a quaint and faintly pejorative term. Nevertheless, it is a vital fact that the teaching at issue here is not the kind that is practiced in back alleys, bowling alleys, kitchens, or colleges, but is the kind that is practiced in elementary and secondary schools. *Portfolio* appears to have been sullied by its use in some recent state initiatives; in the opinion of one informed observer, portfolio procedures "became an exercise in amassing paper" (Olson, 1988). Even so, it seems unwise to have undertaken large-scale use of portfolios for schoolteachers in the virtual absence of small-scale experience (but see Terry and Eade, 1983). In that absence, and in light of the complexity and turbulence of the state initiatives, there remains an open question whether the portfolio idea has been tried and found wanting or has been wanted and found untried (see, for example, Furtwenglar 1985; Teacher Evaluation Study, 1986).

So this exploration will resort not to *nomen-phoria*, but to nostalgia. It will not attempt to untangle an idea of *portfolio* from the complex ventures in which that term has been used recently. Rather, it will attempt to restore the idea of a schoolteacher's portfolio to the undeveloped, problematic, and interesting state in which it formerly lay. The chapter will explore problems associated with borrowing *portfolio* from other occupations, consider purposes that a schoolteacher's portfolio might serve, discuss local arrangements in which portfolios might be constructed, and suggest how portfolios might be fitted to the work of schoolteaching. The chapter will point to various problems without solving them. It will conclude by suggesting that the potential of portfolio procedures

AUTHOR'S NOTE: The author wishes to thank Lee Shulman, Lee Cronbach, Doris Redfield, Howard Gardner, Joan Knapp, and the staff of the Teacher Assessment Project for their contributions to this chapter. Also, to absolve them of any responsibility for its shortcomings. Work on the chapter was supported by the Stanford University School of Education's Teacher Assessment Project, which is supported by the Carnegie Corporation; the chapter does not necessarily reflect the positions either of the Teacher Assessment Project or of the Corporation.

depends as much on the political, organizational, and professional settings in which they are used as on anything about the procedures themselves.

A Borrowed Idea

To begin, notice that "the schoolteacher's portfolio" would not be as interesting if it were called "the teacher's personnel file." The name's attraction lies largely in an analogy: The members of some respected occupations such as architecture carry satchels of their work. The implication is that schoolteachers might be better evaluated — or more valued or respected — if they also carried their work in suitable containers. We should be cautious. The portfolio cannot be plucked neatly from its surroundings in another occupation, plopped down in schoolteaching, and expected to work as it did elsewhere. Like the hasty traveler who takes someone else's suitcase from the airplane's baggage rack, the analogue may end up wishing for underwear that fits.

It may be useful to consider more varied sources. To be sure, painters, photographers (and photographers' models), architects, designers, writers, and professors all have portfolios, or something like them. The contents of these containers tend to be the work's end product, direct evidence of its quality. For these occupations, we would expect the portfolio to contain a relatively small, carefully chosen, and meticulously polished sample of the worker's results. But this is only one way of putting a handle on someone's work.

Pilots, lawyers, and social workers have logs, folders, and files. These containers and their contents are related to the work quite differently than in the previous cases. The contents typically are not products of the work but are records, by-products, or tools. Here we expect not a sampling but a fuller account — all of the pilot's flights, not just the pretty ones. We would expect the contents of such logs, folders, and files to show skill, pride of work, and a degree of self-concern, but we would not expect them to be polished for display any more than we would expect a carpenter to bronze her tools.

Salespersons and distributors are hired and evaluated as much by their customers as by their employers. So along with their resumes and monthly sales reports, they also have catalogs and sample cases that document their claim to deliver the goods. They generally do not claim to have made the goods they sell, but they often do claim some ability to display the goods, explain the goods, select the right goods, or adapt the goods to the customer's needs. We would expect the contents of a sample case or catalog to be complete, often glossy — and a good predictor of the objects that will arrive in the truck or the mail.

In some settings and for some purposes, information about a person's activities and accomplishments is compressed into symbols. When soldiers, scouts, and 4-H club members carry out duties, learn defined skills, or complete prescribed activities adequately or superbly, they receive medals, badges, pins, and ribbons. They tend to accumulate those marks of accomplishment over time, on

tunics and sashes. Apart from the organizations that award these symbols, they are only scraps of metal, cloth, and paper. But if we know the missions, manuals, histories, and norms of the respective organizations, we expect these emblems to show a good deal about their possessors.

All of these containers, their contents, and the persons who hold them tend to take their utility, meaning, and standing from the missions, operations, and lore of the institutions in which they reside. Filling the containers tends to be a normal part of the work; the institution's members draw on shared understandings both to evaluate the containers and their contents and to attribute some identity or standing to the persons whose occupational biographies they partly define. And so it is with *portfolio*.

Yanked up from another field and transplanted into schoolteaching, *portfolio* becomes little more than an entry in a dictionary. It is a dignified name for a mere receptacle or case, but then questions of dignity run close to questions of evaluation. For schoolteaching, a large part of this dignified container's appeal lays in its very vagueness, its disconnection from the current occupation of schoolteaching, and its suggestion that schoolteaching could be better or better regarded than it is. But vague potential is not plausible promise. To become plausible, the borrowed idea of portfolio must be reconstructed for its new setting.

We could start by replacing the image of the art case — or attache, or sample case, or tunic — with another that may be better suited to schoolteaching. By arrangements to be explored here, a schoolteacher would be expected to keep a set of folders or envelopes that document her work. Any of these folders could hold interesting photocopies, photographs, audiotapes or videotapes, clippings from publications, handwritten notes, or official documents. The titles and contents of the folders could change over time.

Now we face some questions: Why would a schoolteacher bother to keep such a thing? What would the documents record, and how would they be used or judged? Who would participate in the work with portfolios? What would these procedures require; what would they be worth, and to whom? Initially, these are not refined questions of procedure or technique. Rather, they are rough problems of purpose, organization, and meaning.

Purposes for Portfolios

Considered as one part of a teacher-evaluation system, portfolio procedures should contribute to sound personnel decisions and to the development of individual schoolteachers. Those processes, in turn, operate within efforts to improve schools and to organize schoolteaching as a profession. Desirably, portfolios also would contribute to those organizational developments.

Personnel Actions and School Reform

Processes for hiring, assigning, tenuring, annual evaluation, and other personnel actions should employ a range of pertinent evidence to reach a fair and valid decision. Portfolio documentation could bring important evidence into the scope of judgment; schoolteachers' plans, materials, assignments, tests, and written feedback to students are natural materials for a portfolio. Notice, however, that schoolteachers are likely to have considerable latitude in constructing their portfolios. When the stakes in evaluation are high, they may be led to polish the documentation to a degree that is out of proportion to its function and importance, or even to misrepresent its source. On the other hand, including portfolios as one part of the evidence for personnel actions might add importance and motivation to portfolio procedures that are intended principally for other purposes such as teacher development. The problem here may be to strike some reasonable balance.

Selecting human resources wisely is one thing; employing those resources effectively is another. Here we move into the terrain of school reform. For example, one of schooling's more peculiar practices is to assign new teachers a full load of teaching with little or no support in some of the department's, school's, or school district's most difficult classes. By itself, a portfolio procedure is quite unlikely to thwart the institutional forces at work. However, in a concerted campaign to treat new teachers more sensibly, routine use of portfolios containing detailed records of new teachers' limited experience could constantly reaffirm the reasons for providing them both reasonable conditions for learning to teach and active support in doing so. If schoolteaching is understood to be complex and difficult work, induction to schoolteaching should be an extended period of support; a portfolio procedure could provide occasions to plan, monitor, support, and record the new teachers' attainments.

Teacher Development and Professional Organization of Teaching

Improvements in preparatory education, induction to schoolteaching, and continuing education for schoolteachers all call for formative assessments, both of persons and of programs. Here, the schoolteacher's portfolio could support continuity in planning, supporting, and monitoring a schoolteacher's professional advance. Portfolio processes might contribute to a more deliberate and cumulative approach both to schoolteaching and to its improvement.

It may be argued with some justification that schoolteaching is among the most complex, subtle, and difficult kinds of work that a person can do. At the same time, it may be suggested that schoolteaching is scarcely cultivated in an organized way. Teachers tend to be unimpressed by their teacher-education programs, tend to work in isolation both from each other and from others who might be useful to them, and tend to work in school conditions that provide scant

encouragement for reflection, mutual assistance, or substantial support (Fieman-Nemser & Floden, 1986; Lanier & Little, 1986; Lortie, 1975; Rosenholtz, 1985). Under these circumstances of difficult work, scarcely cultivated, it is perhaps not surprising either that the core practices of schoolteaching have remained remarkably constant over the last century (Cuban, 1984) or that thoughtful observers are disappointed by common practice (Boyer, 1983; Goodlad, 1984; Sizer, 1984).

In that view of things, the quality of schoolteaching is less a problem of individual merit than a problem of organizing the capacity to help large numbers of ordinary people to do extraordinary work. Without slighting its uses in licensing, hiring, promoting, or firing schoolteachers, one may argue that teacher evaluation's critical function is to provide processes by which schoolteaching can be cultivated in the collective fashion of a art, craft, science, or profession.

Those processes require both a refined language for discussion of practice and norms of professional exchange that sustain a closer and more fruitful examination of schoolteachers' work. Here, portfolio procedures might provide a worthwhile complement both to consultation and to supervision or coaching based on the observation of teaching. Particularly if they are starting from scratch, schoolteachers are likely to find penetrating conversations based on observation of teaching to be difficult and stressful (Little, 1984). They might work up to rigorous talk about their work by looking together at the documents they employ. Such materials are somewhat more distant from the teacher than her own performances in the classroom; there are precedents for such conversations. Likewise, constructing and examining documents may provide focus and depth to consultations formerly based on oral reports.

As schoolteaching is now organized, students' responses tend to be the only substantial contingency on teachers' performance (Rosenholtz, 1985). While schoolteachers often receive vacuous, casual, or uninformed praise (Bridges, 1986), they tend to go without the form of praise that can come only from someone who looks closely at the work and in doing so displays her competence to judge it. In the absence of well-founded evaluative exchanges, schoolteachers are unlikely to be influenced much by each other or by their nominal supervisors; effective professional norms require both opportunity to observe and ability to sanction. The emergence of a tradition of well-founded praise for proficient teaching would be a substantial innovation. A portfolio procedure could provide colleagues something concrete to admire, a discipline for assembling the evidence, an audience to admire it, and an occasion for doing so.

Selection decisions, teacher development, school reform, and professionalization of schoolteaching provide quite enough purposes for any evaluation procedure to serve. The question now is whether portfolio procedures could be arranged to serve any of them.

PARTICIPATION	GUIDANCE		
	Informal Norms	Mixed Sources	Formal Prescriptions
Entries produced mainly by the schoolteacher	ELECTIVE ENTRIES: draft class handouts; a test being written; test results to interpret; the teacher's report on a special project	GUIDED ENTRIES: teacher's written presentation of a lesson plan and materials; an annotated collection of student evaluations	REQUIRED ENTRIES: an application for employment; a questionnaire on participation in continuing education
Entries produced jointly by the schoolteacher and others	COLLEGIAL PRODUCTS: notes on a consultant; a unit planned with a colleague; an exchange of notes on a point of teaching	NEGOTIATED ENTRIES: joint documentation of teaching practices; joint review of student products; professional development plans	PROCTORED ENTRIES: situated performance assessments, e.g., of test construction, questioning, marking student papers
Entries produced mainly by others	COMMENTARY: letters of recommendation; informal observation notes; parents' notes of praise	ATTESTATIONS: structured reports or ratings of observations; surveys of students or parents; record of a structured interview	OFFICIAL RECORDS: diploma, license, certificate, records of employment and service, records of continuing education

Figure 15.1 Nine Modes of Portfolio Making.

How Could Portfolio Procedures Be Organized

An attempt to employ portfolio procedures is also an attempt to form an organization that values those procedures, understands them, and is equipped to carry them out. That organization problem will be reduced first to two questions. Who will participate in constructing portfolios? What will guide them in their work on portfolios?

Nine Ways to Skin a Cat

Figure 15.1 combines responses to the questions of participation and guidance to name and give examples of nine appreciably different ways to generate a portfolio folder of an item of documentation within it. These options suggest two contrasting images of the portfolio. In one, the portfolio is a *workspace*, a place for a schoolteacher to work at improving his practice. Here it may be desirable for the portfolio to contain a sample of the teacher's *worst* work, so that it may be improved. In the other image the portfolio is a *display* assembled for judgment by others. Unless we suppose that a teacher can be persuaded to indict herself in her own portfolio, we should expect the portfolio-as-display to contain a sample of her *best* work. We face a possible trade-off between the portfolio's

purposes. Just how this tension could or should be resolved is not immediately clear.

Portfolio entries produced by the teacher. Three classes of portfolio entries could be produced mainly by the schoolteacher (Figure 15.1, top row). In *elective entries*, a teacher guided by some informal tradition of portfolio-making would decide what material to include in the portfolio and why, for example, a set of class handouts that need to be rewritten. An organization could define a class of *guided entries* by providing written instructions, such as a list of issues to be addressed in documenting a lesson, to assure that schoolteachers compose portfolio contents in a useful form that is germane to some process or decision. More extensive written rules would tend to define a class of *required entries in prescribed forms*, for example, a questionnaire on continuing education activities.

Jointly-Produced Portfolio Entries. In jointly produced portfolio entries, documents would enter the portfolio in the course of some professional or official exchange. *Collegial products* such as a set of notes on a consultation would emerge from advisement and collaboration among schoolteachers and persons who normally work with them. Adding organizational policies to collegial and supervisory activities would produce a class of *guided and negotiated entries*. For example, a mentor teacher might suggest that a new teacher should place a newly written lesson plan in her portfolio to meet a school district's evaluation requirement. Further formalized, this joint activity could produce *proctored entries*. The mentor might go on to say, "Now, take an hour to construct a test for this lesson, and we'll add it to your folder." In effect, the mentor administers a field-based exercise in test construction.

Portfolio Entries Produced by Others. Three classes of portfolio entries might be produced by someone other than the teacher. *Informal commentary* includes letters of recommendation, written evaluations produced under general guidelines, and parents' notes of praise. Adding formalities and related understandings would define a class of *structured attestations*, for example, a structured observation record, a supervisor's confirmation that a new teacher successfully conducted a lesson of a particular type, or a survey of students or parents. Finally, there is the class of *official records*, including diplomas, transcripts, and records of employment or continuing education credit.

Comments on the Options

While acknowledging their potential value for a schoolteacher's personal reflection and development, one may argue that portfolio entries prepared by the teacher alone have limited prospects either for improving teaching or for evaluating schoolteachers. These entries call for little or none of the collegial or supervisory support that a program of teacher development would imply. They provide little or none of the control or corroboration of evidence that may be needed to sustain personnel actions. They do not clearly engage teachers and

others in substantial conversations, and so have limited utility as a tool of organization, either of schools or of the schoolteaching occupation. By and large, these options seem suited to the current — and debilitating — isolation of schoolteachers.

Joint production of entries might help to derive the most formative value from portfolio procedures, to regulate the volume and nature of the portfolio material, and to assure its origins and relevance to any decision making that may ensue. Note, however, the organizational requirements. The options for joint production presume that professional exchange among schoolteachers is both desirable and ordinary, or even obligatory, and that there are places for that exchange in the school schedule and budget. Further, that local norms encourage such exchange. Finally, that there are common understandings and a language of schoolteaching practice, in terms of which fruitful exchange could be conducted. These are not procedural details; they are substantial organizational accomplishments.

The entries produced by others than the teacher tend to provide independent description and judgment that may be important in personnel actions. Some of the entries, particularly the structured attestations, also could be designed to be maximally informative and useful to the schoolteacher. Otherwise, these entries contribute most to a display. Like the joint productions, they tend to have substantial organizational requirements.

Considering the limitations of written guidance, the value of most of the entries appears to depend heavily on the emergence of an informal lore of portfolio making. Especially where portfolios are intended to improve professional and supervisory interaction, it may be desirable to train teachers, mentors, supervisors, and evaluators together so they can begin to share that lore and establish the working relations they will need. Clearly, making and reviewing portfolios calls for sound, informed judgment. It may be particularly important to establish that colleagues, mentors, supervisors, and evaluators are qualified to play their respective parts, both in the eyes of the schoolteachers with whom they work and in the eyes of other parties to whom they are responsible. They might start to achieve that qualification in common training designed for the purpose.

Portfolios must be well worth looking at to justify the effort that they are likely to require. We should turn now to that question.

What Would the Portfolio Show?

Potentially, a wide array of documents — photocopies, photographs, observation reports, videotapes, etc. — could be included in a schoolteachers' portfolio. How would those documents be related to schoolteachers' work? What value would they have in personnel decisions, teacher development, or organizational development? With few exceptions, the persons who will actually construct, review, and evaluate portfolios will encounter those questions not as issues of measurement, but as problems of mutual meaning or significance. The initial

task is to organize, in and around portfolios, a penetrating and useful conversation about schoolteaching. Then measurement technique might add refinement to the procedure.

Fitting the Portfolio to Schoolteaching

As compared to pencil-and-paper tests, assessment-center procedures, and low-inference observation schedules tied to lists of schoolteacher competencies, the main attraction of portfolios appears to be that they might provide opportunities to see a schoolteacher's work relatively whole and in context. In portfolios, we would not regard the teacher as a bundle of traits or competencies, nor would we see schoolteaching as a list of research-based rules. A teacher-evaluation system would have more efficient ways to examine teachers and teaching along those lines, if that is thought to be desirable. Instead, we would be interested in schoolteaching as a form of expression, a humane project, an evolving state of affairs, and a situated accomplishment over time. The portfolio would be the place for schoolteachers to define their situations, to interpret what they see to be important, and to document the practices that they hold to be good for those situations. Thus, a portfolio procedure should provide schoolteachers wide latitude to express their practice.

However, we may suspect that constructing a revealing and useful portfolio is likely to be an achievement in its own right and of its own kind. Educational researchers and philosophers have shown repeatedly that it can be difficult to reveal important aspects of schoolteaching. It would be naive to suppose that, simply because schoolteachers do the work, they would do a better job of revealing it in portfolios. General procedures and generic suggestions about recording media are unlikely to help them much. Most new cameras come with operating instructions, yet the world is littered with snapshots that only their makers can interpret or admire, all pasted into albums that reveal to an observer little more than the month and year in which the film was developed. One doubts that such evidence could be much improved by adding little circles and arrows to the pictures, and paragraphs on the back telling what each one is all about. Without a refined tradition for documenting their work, schoolteachers might be expected to produce portfolios that would bore them and would-be evaluators. Learning how to study and reveal one's own teaching may be one of the most formidable challenges to the portfolio idea.

Schoolteachers may find it somewhat easier to capture important aspects of this activity if they are provided some portfolio structure that suits the nature of their work. Clearly, the complexity of schoolteaching combines with the variety of ways of looking at it to preclude any definitive arrangement or description of portfolio elements. All such arrangements should be regarded primarily as conventions. That is, they should help schoolteachers to choose material for portfolios and to arrange it ways such that they and others who share the conventions can interpret and judge the work that is shown.

Following paragraphs attempt to suggest one structure for portfolios that could have strong ties with the work and help schoolteachers to describe their own practice. Generally, the portfolio would be divided into sections that distinguish clusters of schoolteaching tasks or problems — characteristic situations that call for purposeful action by the teacher. Within a section, the basic element of a portfolio would be a folder or entry that would attempt to show how the schoolteacher addresses or resolves the essential elements of a given problem in the teacher's situation.

Schoolteachers' Activities

Consider five entwined clusters of schoolteachers' activity: teaching a class, planning and preparation, student and program evaluation, interaction with other educators, and interaction with parents and members of a community. Each cluster comprises a distinctive array of problems — situations, events, or matters that call for teacher action. The clusters are not independent realms, but are thoroughly entangled. Teaching a class might be taken to be the central domain; the others are strongly shaped by or oriented to it. The list implies that there would be five broad sections in the portfolio; it also hints at some of the connections we might expect to make among those sections. In an evaluation program, it might be sensible to rate the five sections separately, on the grounds, for example, that new teachers and experienced teachers, or teachers who work in different settings, could be expected to have different profiles of activity.

However, the proposed sections remain far too broad to help schoolteachers prepare their portfolios; "Show how you evaluate students" is a sweeping mandate. In the absence of a basis for choice of documentation, amassing paper would be the likely response. A theoretical inclination would lead us to distinguish analytical elements of teaching, such as management and discipline, command of a subject matter, or attention to student differences. "Show how you take account of student differences in planning and preparation" does help to limit the terrain for a given portfolio entry. However, that move also tends to disintegrate schoolteaching. It suggests that we would prepare a list of discrete teaching competencies, rate the teacher separately on each, and reach a decision or rating essentially by addition. The prospect of seeing the schoolteacher's practice relatively whole and in context fades rapidly.

Therefore, it may be preferable to adopt a vocabulary of schoolteaching problems: "Show how you start your math class at the beginning of the year . . . plan a unit on the American revolution . . . teach a science lesson incorporating a small group activity . . . counsel the parent of a student who is not learning to read . . . mark and grade a set of student essays on *The Pain and the Great One* . . . coach a less experienced colleague through a class debate." The aim would be to relate the analytic elements of schoolteaching problems, to examine how the teacher resolves them in a practical situation, and to judge that resolution more or less as a whole.

TASKS	CONCERNS			
	Schoolteacher's responsibility	Subject matter	Individual students	Class Organization
Teaching a class	modelling curiosity or courtesy; justifying a task, procedure, or rule	showing how to write a simple sentence; presenting elements of a theory	checking a student's understanding, giving help	teaching a class rule or routine; maintaining the routine in a given case
Planning, preparation	setting academic and social goals for a course or unit	selecting and sequencing topics for a lesson	adapting audio-visuals or hand-outs for students	selecting or designing a small group procedure
Student evaluation	using evaluation results to redesign instruction	composing a question, task, test or assignment	making comments on a student's product	assessing a class's progress toward civility, cooperation
Professional exchange	debating aims and means of "discipline" at a faculty meeting	asking a colleague for current information on a topic	talking with a school counselor about a student	coaching a colleague through a unit that uses a class debate
Community exchange	explaining school goals at back-to-school night	recruiting a guest speaker; soliciting materials	meeting with a parent or social worker	leading a field trip to a museum, park, or workplace

Figure 15.2 Some Activities That Might Show in a Portfolio

Fourfold Problems of Schoolteaching

So, continue the convention-building by considering (after Schwab, 1978) that most problems of schoolteaching involve four entwined concerns that must be resolved more or less simultaneously. First, schoolteachers are responsible to students and to others for an array of important principles and results including protection of children, promotion of citizenship, pursuit of justice, and mastery of school subjects. That partial list cannot sensibly or responsibly be reduced to its last item; there is no simple bottom line for schoolteaching.

Second, the teacher conveys or represents a subject matter to students, or engages them with it. That subject matter is not just a collection of information and skills, but also is a structure of perspectives and theories, a set of ways of knowing and learning, and an organized field of work in the world. To teach the subject matter is to somehow give students access to, and to help them gain a grip on, all those features of it.

Third, the teacher works with and in a crowd of students who must be considered as individuals. They are not blank slates or unmolded clay, but are experienced actors in the situation. They are not identical, but are differentially equipped and differently inclined; they hold different memberships in different settings outside the class and outside the school. Nothing that the teacher can do will affect them all in the same way.

Finally, the crowd of students cannot be understood only as a collection of individuals. They constitute an actual or potential organization that both has its

own dynamics and is affected by other social units — cliques in the classroom as extensions of the student or neighborhood society. The class organization has its own distinctive properties, problems, and educative or miseducative influences. It is both a means for instruction and an element of curriculum.

In paragraphs, those four concerns may be separated. In practice, they are intricately entangled. Resolving them simultaneously in one coherent line of action is a schoolteacher's chief accomplishment. Arguably, the central purpose of a schoolteacher's portfolio entry is to document whether and how a teacher achieves that complex resolution for a given problem in a given situation, for example, starting a math class at the beginning of the year. This is the sense in which a portfolio entry should show the work whole and in context.

Activities That Might Show in a Portfolio

In resolving their fourfold problems, schoolteachers engage in a variety of activities that may lend themselves to documentation. Figure 15.2 combines the foregoing discussions of schoolteachers' activities and their problematic elements to display a range of more specific actions that might be desirable to show in a portfolio. (Reading across rows helps to reduce this display's tendency to disintegrate schoolteaching practice.) The question now arises how schoolteachers would show those activities, either to themselves or to their evaluators.

Various techniques that have been mentioned above might be used to record, document, and describe schoolteachers' activities; their respective virtues should be explored at a length that is not possible here. Note that economy and convenience are likely to dictate the possible techniques in many instances. If we keep in mind that portfolios might include material produced jointly by the schoolteacher and a colleague or by the colleague, the range of possibilities may be expanded. It may be reasonable to answer adapting techniques used in other sorts of teacher evaluation and in research on teaching.

As we consider techniques, we also should ask whether any documentation procedure is likely to serve well in the absence of prior agreements and common understandings about the ways in which portfolio documentation will be understood and used. For example, one large source of potential documentation is the material that schoolteachers already do — or advantageously could — produce in the normal course of their work. Portfolio-making might emerge as part of an occupational tradition of saving, sharing, and refining schoolteachers' solutions to the problems they face. Often, this will be rough material, working documents that will look right in the portfolio-as-workspace but may seem out of place in the portfolio-as-display. How would a portfolio reviewer regard documentation of a flawed piece of teaching that is accompanied by evidence that the teacher detected the flaws and has ideas about how to correct them? Is this a sorry teacher who made a mistake, was forced to recover from it, and is now trying to put the best face on things, or an admirable teacher who reflects on her work and improves steadily? Of course it depends; the point is that the teacher who makes

the portfolio will need to know not only what the evaluator thinks it depends on, but also what stance that evaluator is taking toward the purposes of teacher evaluation. Such understandings are not readily produced by printing procedures manuals. The organization of schools and the politics of school reform are very much in play here.

A Moving Focus of Attention

Even if it were feasible, documenting any large proportion of a schoolteacher's practice at one time would tend to produce a portfolio that is both very thick, in terms of the total amount of material it contains, and very thin, in terms of its ability to shed light on the schoolteacher's resolution of any particular problem. Formative operations provide one rationale for focusing attention: The material in the portfolio should concern those schoolteaching problems that a teacher at a given time is learning to solve or to solve better. Summative procedures might take a similar tack; there may be parallels here both to cycles of clinical supervision and to annual cycles of professional development planning. In a given year, a few problems of schoolteaching would be selected for detailed examination in the portfolio. If the teacher's performance is found to be unsatisfactory, there would be a substantial basis for teacher development. If the teacher's performance is found to be satisfactory, then all the portfolio material could be replaced by highly condensed notes to that effect, perhaps in the teacher's personnel file.

So far, the discussion has concentrated on teaching practices. What about the cognitive, affective, and social outcomes for students? It is debatable whether student achievement can be attributed to teacher activity in a manner that is usable in a teacher's evaluation (see Glass, this volume; Haertel, 1986; Millman, 1981; Redfield, 1988). It must be even more debatable whether the problems can be resolved in workable technology for any schoolteacher who constructs a portfolio. Linking goals other than student achievement to schoolteachers' activities or characteristics is unlikely to be easier. Even so, portfolios could address the questions of outcomes. Materials describing practice could shed light on the teacher's conception of the goals to be pursued in teaching, his means for obtaining some estimate of the results attained in his classes, and his uses of the information he obtains. Focusing on a few problems of schoolteaching at a time could make it more feasible to go into the question of outcomes from that angle.

As a form of accountability, a multiyear year cycle of close examinations of selected parts or problems of a schoolteacher's work would appear to be at least as stringent as annual attempts to rate the teacher on the whole of the teacher repertoire. More importantly, perhaps, the moving focus of attention could make it more feasible to *attain* the standard that the evaluation is attempting to *set*.

Conclusion

Consider "the schoolteacher's portfolio" as a potential tradition that incorporates a refined lore of portfolio making. At any time, a schoolteacher would be expected to have a set of folders that show how the teacher handles some of the fourfold problems that he faces in schoolteaching. The set of problems to be addressed in a given sequence or cycle of assessment could be specified by an organization, negotiated by the teacher and some official or colleague, chosen by the teacher, or all three. Early in a sequence, the material in many of the folders would be sketchy. Over time, colleagues and supervisors would work with the teacher to examine her work, to improve it, and to judge which materials could or should be replaced, either by doing the job better or by making a more useful and revealing record. In many instances, the colleagues or supervisors would help the teacher to produce — or would themselves produce — material for the folders. Through successive cycles of assessment and development, the teacher with support would address many of the complex problems that she faces. So far, the portfolio is mostly a workspace, and tool for the development.

From time to time, the schoolteacher would present a sample of her current or past portfolio folders as part of an annual evaluation, a tenure review, or other personnel action. When arranged as a display for these purposes, the portfolio would be representative, not of all of the teacher's work, but of the teacher's best work. The sample would not be comprehensive, but would be diversified. At no time would a teacher or an evaluator pretend to be able to make a comprehensive and thorough examination of a schoolteacher's practice. Over time, the teacher, with assistance from others, would examine many aspects of her practice and would present that work for evaluation.

That image of the portfolio suggests a larger portrait of the schoolteaching occupation. Schoolteachers would spend considerable time working with their portfolios and with their colleagues to examine, refine, and share a growing stock of strategies, practices, plans, activities, and materials — a body of increasingly refined solutions to the concrete problems of schoolteaching. Schoolteaching would be a more deliberative and cumulative practice than current conditions often permit or encourage. Formative and developmental processes would bear the main weight of attaining high standards in schoolteaching. Summative evaluation procedures would have the secondary functions of lending some direction and motivation to the formative activity, providing institutional recognition for good work and providing assurances to school boards and the public.

That portrait clashes with contemporary patterns of school organization. As sketched here, portfolios do not fit well in the current scene. In the present rush to reward, fix, or fire schoolteachers by essentially bureaucratic means, portfolios would be made to bear considerable summative weight — in the absence either of effective formative support or improvement or of a refined tradition that helps schoolteachers to make useful and insightful examination of their work. Under those conditions, portfolios could be expected to degenerate from instru-

ments of improvement into mere displays, "an exercise in amassing paper," and a waste of time both for the teachers who prepare them and for the persons who review them.

It remains to be seen whether, in any conditions, the schoolteacher's portfolio can be useful either for schoolteachers or for their evaluators. The idea appears plausible enough to merit development. From the point of view of pacing that development, it may be an advantage if "portfolio" has ceased to be interesting for large-scale initiatives. Well-founded local efforts to improve schools as workplaces, to improve continuing education for schoolteachers, or to strengthen their professional interaction could be good settings for developing and testing variations on the portfolio idea.

The author's stance is that the schoolteacher's portfolio belongs to an occupation of schoolteaching that we don't have now, but should be organizing for the future. In that occupation the portfolio would be more a means for schoolteachers and others to cooperate in attaining a high standard of practice than a procedure for evaluating them. The portfolio would derive its means and standing more from the lore of schoolteachers than from the lore of schoolteacher evaluators or evaluation specialists. The productivity of portfolios would be more a question of policy, school organization, professional relations, and professional discourse than a question of measurement technique.

References

- Boyer, E.L. (1983). *High school: A report on secondary education in America*. New York: Harper and Row.
- Bridges, E.M. (1986). *The incompetent teacher*. Philadelphia: The Falmer Press.
- Cuban, L. (1984). *How teachers taught: Constancy and change in American classrooms 1890-1980*. New York: Longman.
- Fieman-Nemser, S. & Floden, R.E. (1986). The cultures of teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching*, (3rd ed.), (pp. 505-526). New York: Macmillan.
- Furtwenglar, Carol. B. (1985). *Evaluation procedures in the Tennessee career ladder plan*. A paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Glass, [chapter 14 in this volume]
- Goodlad, J.I. (1984). *A place called school: Prospects for the future*. Highstown, NJ: McGraw-Hill.
- Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational analysis and policy evaluation*, 8(1), Spring, pp. 45-60.
- Lanier, J.E. & Little, J.W. (1986). Research on teacher education. In M.C. Wittrock (Ed.), *Handbook of research on teaching*, (3rd ed.), (pp. 527-569). New York: Macmillan.
- Little, J.W. (1984). *Professional development roles and relationships: Principles and skills of "advising"*. San Francisco: Far West Laboratory for Educational Research and Development.
- Lortie, D.C. (1975). *Schoolteacher: A sociological study*. Chicago: University of Chicago Press.
- Millman, J. (1981). Student achievement as a measure of teacher competence. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 146-166). Beverly Hills, CA: Sage Publications.
- Olson, L. (1988, June 8). Capturing teaching's essence: Stanford team tests new methods. *Education Week*, 1, 20.
- Redfield, D.L. (1988). Expected student achievement and the evaluation of teaching. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Rosenholtz, S.J. (1985). Effective schools: Interpreting the evidence. *American Journal of Education*, (May), 352-387.

- Schwab, J.J. (1978). The practical: Translation into curriculum. In I. Westbury & N.J. Wilkof (Eds.), *Science, curriculum, and liberal education: Selected essays* (pp. 365-383). Chicago: University of Chicago Press.
- Sizer, T.R. (1984). *Horace's compromise: The dilemma of the American high school*. Boston: Houghton Mifflin.
- Teacher Evaluation Study (1986). *Beginning teacher portfolio study*. Tampa, FL: University of South Florida. (ERIC Document Reproduction Service No. ED 277 689.)
- Terry, G.L. & Eade, G.E. (1983). *The portfolio process: New roles for meeting challenges in professional development*. A paper presented at the annual conference of the Association of Teacher Educators, Pensacola, FL. (ERIC Document Reproduction Service No. ED 229 342).

16. Conventional Tests for Licensure

George Madaus

William A. Mehrens

The determination of who will receive a license to teach is a matter of great importance — both to the 5% of the population who become teachers and to the 100% who become students (Pratt, 1977). Historically, the United States has gone from giving licensing tests to just relying on graduation from an accredited college, back to requiring licensure tests (Haney, Madaus, & Kreitzer, 1988). Currently 44 states either test prospective teachers as a licensure requirement or plan to implement such a requirement in the near future (Rudner, 1988). Of those 44 states, 31 test for basic skills, 24 for professional knowledge, and 29 for subject area knowledge.

The motivating factor behind the current popularity of teacher-licensure tests is that the public and many educators believe that both our colleges and our state licensing boards have failed as gatekeepers and that incompetents are obtaining licenses to teach. There is considerable debate about the correctness of this perception (Haney, Madaus, & Kreitzer, 1988; Mehrens, 1987a, 1987b). Haney, Madaus, and Kreitzer reviewed various estimates concerning the number of incompetent teachers and found it to be about 10% of the current teaching staff. These estimates are about the same estimates one finds for other professions such as law or medicine.

Whether tests are the best way to ensure that teachers have the minimum skills and knowledges to be licensed (or to be considered competent) is also a debatable point. Other chapters cover such topics as preservice evaluation (chapter 5), classroom observations (chapter 13), and performance tests (chapter 19). These chapters detail useful and necessary evaluation procedures. However, the focus of this chapter is on conventional (i.e., paper-and-pencil, for the most part

AUTHOR'S NOTE: Acknowledgments are gratefully extended to Steve Murray and Larry Rudner who provided detailed and helpful comments on both the outline and first draft of this chapter. We would also like to thank our colleagues at Boston College and Michigan State University for helpful comments. These colleagues are Amelia Kreitzer at Boston College and Robert Craig, Irvin Lehmann, Susan Phillips, and Herbert Rudman at Michigan State University.

The order of the authors is alphabetical. They have shared equally in the writing task.

multiple-choice type) tests, which must be passed as a necessary (but not necessarily sufficient) condition before obtaining a license to teach.

This chapter is aimed at those who must build, who commission to have built, or who use paper-and-pencil tests for teacher licensing. While we believe that the traditional paper-and-pencil test will always have an important place in teacher licensing, we anticipate that its future use may be more as a first step in a comprehensive teacher-assessment program than is currently the case, for two reasons. First, the current work on teacher assessment for board certification of Shulman and his colleagues at Stanford has revealed the popularity of assessing skills not easily handled by the traditional methodology of multiple-choice examinations. Second, we anticipate the most widely used traditional teacher-licensing test (the NTE) will be replaced by a new battery that will contain a conventional test but be broader in scope and more compatible with the approach of the National Board's assessment approach. (We do recognize that board certification and licensure have two somewhat different goals.) As Zieky has pointed out, the problem in using performance tests for teacher certification may eventually be overcome. However, until that time, we should continue to make the objective tests as fair, as reliable, and as valid as possible (Zieky, 1988).

Purposes of Licensure

The United States Department of Health, Education, and Welfare defined licensure as follows:

Licensure: The process by which an agency of government grants permission to persons to engage in a given profession or occupation by certifying that those licensed have attained the minimal degree of competency necessary to ensure that the public health, safety and welfare will be reasonably well protected (1971, p. 7).

Note that licensure is to protect the public from those likely to do harm. It is a selecting-out process. This has important implications for test validation. The inference made from a test score is that a person falling below the cut score lacks the minimum skills, knowledge or prerequisites necessary to ensure the public welfare. Notice that there is no inference about being successful — only about not being minimally competent. The degree to which this inference is open to sound empirical investigation is a subject of considerable debate. The authors return to this point below.

Licensure procedures are to determine whether or not individuals have *minimal* competence. This is a quite different purpose from that of employment tests. Employment tests are intended to help identify those applicants for a job who are likely to be the most successful. The AERA/APA/NCME *Standards for Educational and Psychological Testing* (hereafter called the *Standards*) clarify the differences in a succinct manner:

For licensure or certification the focus of test standards is on levels of knowledge and skills necessary to assure the public that a person is competent to practice, whereas an employer may use tests in order to maximize productivity (1985, p. 63).

Advantages of Conventional Tests

The most commonly used item format for licensure examinations is the multiple-choice item (Shimberg, 1982). This seems quite appropriate because the purpose of most licensure tests is to see whether or not the applicants have the necessary knowledge. Knowledge is a particularly important aspect of teacher competency. As Shulman has pointed out: "Teaching is, at its root, a learned profession. A teacher is a member of a scholarly community" (n. d., p. 12).

Almost all authors of measurement texts have advocated the use of multiple-choice items for measuring knowledge (see for example, Bloom, Madaus, & Hastings, 1981; Mehrens & Lehmann, 1984; Nitko, 1983). There is a wide body of literature demonstrating that multiple-choice items can measure knowledge. Conventional tests can adequately sample a broad domain helping ensure that some necessary knowledge is not left untested. Conventional tests are efficient to administer and score.

Limitations of Conventional Tests

Pottinger (1979) has argued against conventional multiple-choice tests because they do not do an adequate job of protecting the public. That is, they let too many incompetent people get certified. Generating a solution to a problem is more difficult than the multiple-choice test question task of choosing a solution. Pottinger argued that the real task facing teachers may be the former. Thus, teachers may get certified who have enough knowledge or skill to choose a correct solution on a multiple-choice question but not have enough competence to generate a solution.

Pottinger is not alone in his criticism of multiple-choice tests. Others have argued that multiple-choice tests keep competent people out. Such critics base their criticism on the notion that some people are unable to adequately demonstrate their knowledge on multiple-choice tests. They claim that the format may produce items irrelevant to competence in the classroom.

Still another argument is that multiple-choice tests do not measure a host of other characteristics such as love, compassion, and dedication, which are important for a teacher. We dismiss this argument out of hand. It is not that we don't believe that such traits are essential to teaching, but that teacher tests make no claim to measure them. This is not a criticism of the tests. Just as it would be foolish to criticize a bathroom scale because it does not measure something it

was not intended to measure. In fact, those who know anything about discriminant validation would be dismayed if a test designed to measure knowledge actually measured love. It would be an indication of a lack of validity! Instead, teacher tests measure other important traits also important to teaching.

The authors agree that passing a multiple-choice test does not ensure that one will be a good teacher — or necessarily even a minimally competent one. The authors do not suggest that passing a conventional licensure test be the only criterion for licensure. Indeed one should use test data in conjunction with other information about a prospective teacher. The data from all sources should be combined in some fashion to make the single decision about licensure (see chapter 20).

Inferences from Teacher Licensure Tests

Prior to discussing test construction procedures, just what inferences are to be made from a licensure test need to be specified. Further, it seems important to distinguish the inferences test builders and test users wish to make from the inferences that others may draw (or claim one cannot draw) from the scores. For a list of the kind of inferences policy makers, the media, the courts, state departments, and test contractors make from teacher tests, see Madaus (in press).

The builders and users of tests have a responsibility to gather evidence and logic to support their particular inferences. In the process of doing this they may use logic and/or evidence to rule out the plausibility of some potentially competing inferences. Thus, it becomes important to carefully analyze the inferences the users intend to make from teacher-licensure tests in order to gather the proper mix of validity evidence to support them.

Further, users have the responsibility to warn others against drawing unwarranted inferences from the data. However, they do not have the responsibility to gather evidence to support (or refute) all inferences others may make (or claim can not be made) from the test scores. This point needs to be stressed because a common method of attacking the use of tests is to state that there is no evidence that the scores measure some variable that the users/builders never intended the scores to measure.

Of course, if test builder/users do not wish others to impute incorrect uses from the scores, they have a responsibility to make clear just what inferences they wish to draw, and the evidence or logic supportive of those inferences. The *Standards* state that for licensure tests the

... focus of test standards is on levels of knowledge and skills necessary to assure the public that a person is competent to practice (AERA/APA/NCME, 1985, p. 63).

The meaning of the word *competent* in the above quote is subject to various interpretations. However, having the competence to do something does not mean

a person will do it or will do it well. Thus, there is a difference between competence and performance. Further, performing in a certain way does not guarantee any certain outcome, so there is a difference between performance and effectiveness. Medley (1982) makes some clear distinctions between teacher competency, competence, performance, and effectiveness. He defines the four terms as follows:

Teacher competency: Any single knowledge, skill, or professional value position, the possession of which is believed to be relevant to the successful practice of teaching. Competencies refer to specific things that teachers know, do, or believe but not to the effects of these attributes on others.

Teacher competence: The repertoire of competencies a teacher possesses. Overall competence is a matter of the degree to which a teacher has mastered a set of individual competencies, some of which are more critical to a judgment of overall competence than others.

Teacher performance: What the teacher does on the job rather than what she or he can do. Teacher performance is specific to the job situation; it depends on the competence of the teacher, the context in which the teacher works, and the teacher's ability to apply his or her competencies at any given point in time.

Teacher effectiveness: The effect that the teacher's performance has on pupils. Teacher effectiveness depends not only on competence and performance, but also on the responses pupils make. Just as competence cannot predict performance under different situations, teacher performance cannot predict outcomes under different situations.

Because a licensure test is to ensure minimal competency, many of the licensure tests have been called *competency tests*. Generally, the definitions and stated purposes of licensure tests are in accord with the definition Medley used for teacher competence.

There is some agreement about what one can and cannot infer from most currently used teacher-licensure tests. Because they sample a domain of knowledge that is judged by a representative sample of educators to be relevant, one can make inferences about the degree to which an applicant knows a domain of knowledge that fellow professionals consider to be important. This is not a trivial inference. The value of this inference should become clearer when the authors discuss just how the domain was determined in a later section. However, many would argue that support for the inference that a person lacks the minimum skills or knowledge to be successful in the classroom cannot rest exclusively on the judgments of what fellow professionals believe or consider to be important (Madaus & Pullin, 1987).

Because the decision regarding the right to practice in the profession is made in part from the score on a licensure test, the indirect inference, as mentioned earlier, is that a person who has less than some defined minimum knowledge and skills has an insufficient amount of the knowledge and skills necessary to be

minimally successful in the classroom (Madaus wording) or to ensure the public welfare (Mehrens wording). Even if the correct domain of knowledge and skills is measured accurately, where one sets the cut score will influence the truth of the inference made. Thus, the issue of where to set the cut score can be considered an important subset of the validity question. (Standard setting is covered in chapter 21.)

There is little (if any) direct evidence which would allow one to infer anything about degrees of performance or effectiveness. As Medley pointed out, performance and effectiveness are different from competence. However, a host of studies (see Hunter, 1986) do show that job knowledge is correlated with job performance. Further, research shows that in many cases one can generalize from a large set of validity studies to what would likely be true in other situations (validity generalization — see Hunter). Thus, it is reasonable to assume that in teaching, as in all other fields studied, job knowledge will correlate with job performance. Many would argue that it is especially true when one is measuring knowledge of the subject matter that the teacher is to teach. The saying that “one cannot teach who one does not know” strikes many as being quite logical.

Nevertheless, because of the lack of direct evidence the authors caution against making performance inferences. Certainly the correlation between job knowledge and performance is far from perfect. In fact, a review of the literature reveals that:

... teacher tests, have little, if any, power to predict how well people perform as teachers, whether that performance is judged by ratings, of college supervisory personnel, ratings by principals, student ratings, or achievement gains made by students taught (Haney, Madaus, & Kreitzer, 1988, p. 199).

Of course many (e.g., Mehrens) believe the low correlation with *ratings* is because of the criterion measurement problem, which is especially severe in education. *Achievement gains* have some real limitations as criteria also and many would deplore their use as a measure of teacher effectiveness (see Glass, chapter 16 this volume; Porter, 1988, p. 12).

A more recent study of the relationship between subject matter knowledge measured by the Georgia Teacher Certification Tests (TCT) and teaching skills — the later measured by a carefully constructed Teacher Performance Assessment Instrument (TPAI) developed by the state of Georgia — indicated the two were not correlated significantly. In fact, the “Understands subject” component of the TPAI only correlated 0.04 with the TCT subject knowledge test (Guyton & Farokhi, 1987). Madaus believes that what is interesting here is that two measures of subject matter knowledge both claiming to be valid have little convergent validity. He suggests that it is too facile to dismiss the TPAI as subjective and accept the TCT scores as objective given the care with which the TPAI was developed and administered. Mehrens would dismiss the relevance of the 0.04 correlation. “Demonstrates command of school subject being taught” is

only *one* of 30 indicators on the TPAI. Observations are made only on *two* different occasions. The reliability and validity of that single indicator are open to question. Certainly there is at *best* a *very* small sample of content knowledge being evaluated. Restriction in range of scores on both measures is a problem. Finally, almost all individuals pass the TCT, indicating they do have sufficient subject matter knowledge. Correlations of scores *above* the cut score to *any* performance are irrelevant to the validity of the test, as was pointed out early in the chapter.

Test Development Procedures

There is considerable agreement about what general procedures one should employ when correcting a teacher-licensure test. However, there is some debate about just how to conduct the specific steps in a test development process and how well those steps need to be executed in order to produce a quality product. The procedures that will be discussed are as follows: 1) outline and objectives (competencies) development; (2) job analysis survey; (3) domain specifications; and (4) item writing and review.

Outline and Objectives (Competencies) Development

In developing the original outline and objectives one must understand the purpose of a licensure examination. As discussed, the purpose of such an examination is to determine whether the applicants have the level of "knowledge and skills necessary to assure the public that a person is competent to practice" (AERA/APA/NCME, 1985, p. 63).

The necessary first step for developing the outline and objectives is to consider who to appoint to do this task. It is generally agreed that this task should be accomplished by a committee and that the individuals should be experts within the field. What constitutes "experts" is a subject of some disagreement, but most writers believe the committee should include but not be limited to practicing teachers (Madaus & Pullin, 1987; Mehrens, 1987a, 1987b). The committee should include supervisors, university professors, and/or state department personnel. The majority should probably be practicing teachers. All members should have credibility with the appropriate constituents. The criteria to be used in appointing the committee members are subject to some debate. It would certainly be wise to have geographical and gender representation on the committee. Given issues of equity, and the known negative impact of these tests on minority populations, minorities should be adequately represented. The teachers should probably have a minimum number of years of experience — although the authors are not willing to set a particular number.

The starting point for the committee should be a thorough review of the relevant literature (Burns, 1985). What kind of knowledges and skills should be

considered critical? The development of the outline and objectives should include a thorough review of the competencies tested in other states, the scope and content outlines from state departments of education, and the literature on teaching effectiveness. Note that this is not the same thing as trying to establish the "curricular validity" of an examination. The purpose of going to the literature is to find out what is critical, not to find out the subset of content that is taught in all curriculums. Remember that licensure exams have gained in popularity partly because many perceive that some institutions do not turn out competent graduates.

An additional literature source is the research on how teachers spend their time in the classroom. Rosenfeld, Thornton, and Skurnik (1986) reported that they found the categories used by researchers on classroom observations to code teacher activities useful in their job analysis.

The literature review would be somewhat different for examinations in pedagogy than for examinations in subject matter fields. To determine the appropriate content of subject matter tests it would be appropriate to search the curricular materials in the appropriate grade levels for which licensure will be given. However, the authors do *not* suggest that teachers only need to know the subject matter taught in courses or grades in which they are certified to teach. As Shulman has pointed out:

We expect that the subject matter content understanding of the teacher be at least equal to that of his or her lay colleague, the mere subject matter major. The teacher need not only understand *that* something is so; the teacher must further understand *why* it is so, on what grounds its warrant can be asserted, and under what circumstances our belief in its justification can be weakened and even denied. Moreover, we expect the teacher to understand why a given topic is particularly central to a discipline whereas another may be somewhat peripheral. This will be important in subsequent pedagogical judgments regarding relative curricular emphasis (1986, p. 9).

Another approach occasionally taken in addition to the literature review is to interview teachers. This was done by the ETS team (Rosenfeld, Thornton, & Skurnik, 1986) and by the University of South Florida in their construction of licensure examinations for the state (Lavelly, 1987).

Job Analysis Survey

Licensure examinations must be job relevant or job related. The *Uniform Guidelines* state that

Any method of job analysis may be used if it provides information for the specific validation strategy used (Equal Employment Opportunity Commission, 1978, p. 38300).

The *Guidelines* do suggest that the analysis must yield an operational definition of the domain in that the content of the domain should be necessary for critical or important work behaviors. Two commonly accepted methods of determining job relatedness are through document review and group discussion. These two methods (commonly called *logical analyses*) should be employed by the committee developing the list of original competencies discussed in the previous section. This process of using the pooled judgment of a committee of experts is a recognized approach to determining job relatedness (APA, 1980). One major advantage of such a logical analysis is the use it makes of the body of existing knowledge (Kane, Kingsbury, Colton, & Estes, 1986, p. 1.6). However, it may overlook important aspects of the job.

Another common approach to job analyses is observation. While some advocate extensive observations as a component of any job analysis for teacher-licensure tests, others do not view that as a particularly efficient or even useful way to go because they argue that the existing literature already reports the results of teacher observations and reinventing the wheel should not be necessary.

While it may not be considered absolutely necessary, most who build licensure tests conduct a job analysis survey to confirm, disconfirm or add to the judgments of the committee experts (Pechione, Tomala, & Forgione, 1986; Yalow & Collins, 1985). Just how to conduct this job analysis survey is a subject of considerable discussion. Final agreement upon all answers may never be forthcoming, but some of the issues can be discussed. The more salient issues seem to be as follows: (1) How should the tasks (objectives, competencies) in the survey be worded? (2) How should the questions to the respondents be worded and what scale should they be given on which to respond? (3) Who should receive the survey and how many respondents are needed to adequately represent the intended population of respondents? (4) How should the data be analyzed? And (5) How binding should the survey results be on the committee that builds the test? These are difficult questions and unfortunately there is no agreed-upon best answer to them. Nonetheless, a carefully design and executed survey can supply another important strand of validity evidence.

Wording of the Tasks. The job analysis survey should contain a list of the tasks, objectives, or competencies that (based on the literature review and the judgment of the committee) have some reason to be considered important or necessary skills or knowledges important for minimum competence on the job. The words tasks, objectives, and competencies do not connote exactly the same thing to all measurement specialists. Because the inference in licensure (as opposed to employment) decisions is on minimum competence as opposed to predicting productivity (performance) on the job, the general consensus is that the questions should be focused on the knowledge necessary to be minimally competent rather than on the actual work done on the job. (See the AERA/APA/NCME *Standards*, 1985, p. 63). Of course not everyone agrees with this. Some think that such a focus means one has not analyzed the job adequately (Madaus, in press).

Another aspect of the wording issue that is not resolved is how specifically the competencies should be stated. If questions are stated too specifically, they easily can be considered not necessary, because no single bit of knowledge is probably absolutely necessary to teach. However, if they are stated too broadly, they too easily can be considered necessary.

Task Given the Respondents. For employment tests the job analysis surveys ask respondents to rate the importance and/or frequency of the task. Often considerable emphasis is placed on the frequency aspect. For licensure tests, where the goal is to protect the public, less emphasis should probably be given to the frequency rating. In a recent ETS job analysis, a time-spent rating scale was included in the original draft of the inventory but was eliminated based on advice from an external advisory committee and the project staff (Rosenfeld, Thornton, & Skurnik, 1986).

The distinction between the meanings of important, crucial, and necessary result in some debate about just what words are to be used in a job analysis survey. As noted earlier the *Standards* suggest that the focus should be on *necessary* knowledge and skills to assure that the person is competent to practice. The *Uniform Guidelines* use the wording *a necessary prerequisite* (EEOC, 1978, p. 38302). Yalow & Collins have argued that indeed the domain tests should be "necessary, not merely desirable" (1985, pp. 5-6).

Mehrens (1987a) is concerned that the use of the word *necessary* could establish an impossibly high standard — especially if one is rating fairly specific competencies. Obviously, there are different opinions with respect to the wording. Madaus and Pullin (1987) submit that establishing mean ratings on a scale such as (1) crucial, (2) important, (3) questionable, or (4) not relevant to minimal success, and then accepting those items whose ratings are above the midpoint is a procedure that is "biased in favor of positive, rather than disconfirming, answers to questions about the knowledge, skills, and competencies needed to teach at some minimal level of effectiveness" (p. 36). While their concern is certainly not limited to the scale points or the analysis of such data, and while their specific comments are with respect to rating items as opposed to statements, it is clear they are not happy with the scale points.

What is obviously needed is carefully designed research on how respondents react to different wording of scales used in both job-analysis surveys and content validation. Ideally, the results of both a job analysis and a content validation should not vary simply because of the vicissitudes of scale wording. Both concepts should be more robust.

Sampling of Expert Opinion. Not much research has been done on who should be sampled by the survey. Generally, the sampling has been on the domain of practicing teachers in the state who are licensed in the field for which the test is designed. Whether the sample should be delimited to "superior" teachers is an issue of some debate — as is the issue of how one would determine which teachers are superior. Another point of contention revolves around the exclusive reliance on the judgment of teachers about what is considered neces-

sary to be minimally competent. Skills or knowledge that practitioners believe or consider necessary for successful job performance can prove ephemeral when put to the empirical test, particularly when the identified domain is measured using the multiple-choice format (Madaus & Pullin, 1987). Some would advocate including teacher supervisors in the sample. Others would wish to include college professors. It is the authors' opinion that both of the latter two groups should be included in the original committee that writes the tasks that go on the survey instrument, but that they do not necessarily have to be included in the job analysis survey.

The whole area of whether one can generalize from the job analysis findings in one state to another is an area that has not been sufficiently investigated. Certainly for an exam to be used for national registry, the assumption would be that the tasks of a teacher in one state are similar to the tasks of a teacher in another state. This seems like a reasonable assumption.

The issue of whether one can aggregate the results from teachers in different grade levels is also an issue that is deserving of more research. Of course if the license gives one the right to teach across a variety of grades, then the test must cover necessary knowledge and skills in all those grades. However, few, if any, licenses would cover all grades from grades K to 12 and it is still an open question whether there are a set of professional knowledge (pedagogy) areas that are necessary both to be minimally qualified to teach in the early elementary school and in high school. While limited space prohibits the discussion of sampling theory, it should be obvious to all readers that both the size and the representativeness of the sample are important issues.

Data Analysis. Several issues are unsettled regarding the data analysis to be done on the job-analysis surveys. (1) If data are gathered on more than one dimension (e.g., frequency and criticality), should they be combined and, if so, how? (2) Should there be cut scores set in advance with respect to how high the rating needs to be before the task is considered relevant? If so, how and where should that cut score be set? If not, how does some committee eventually decide which tasks have been validated? (3) Should the data be analyzed and reported separately by ethnic group (probably) and, if so, should the minority group data alone be sufficient to throw out a task? Space does not permit deliberation on these important questions. While these co-authors have some differing beliefs about the answers to the above questions, they totally agree that more thought and research on the questions would be of value.

Data Binding on the Committee? The general belief by licensure experts is that the final decision regarding whether tasks are deemed valid should be left to the committee members. They should be given guidelines and, further, should be forced to defend any decision they make that violate the guidelines. While the data should influence the ultimate decisions, they should not absolutely force the decisions.

Domain Specifications

As several researchers have pointed out, there is little to guide test developers in translating task-analysis data to domain specifications or test blueprints (Elliot & Nelson, 1984; Schmeiser, 1987, 1988; Schmeiser & Estes, 1986). This is not a situation unique to licensure tests. Measurement experts have never totally agreed on how explicitly the content domain needs to be defined or what algorithms to use to weight the subcategories of the domain. However the domain is determined, it is a judgmental process and "the defense must be prepared to show that the domain is relevant and that weight is properly distributed over it" (Cronbach, 1980, p. 105).

It is generally agreed that for licensure tests the follow three points should be considered: (1) The domain should be fairly broad because a license is for a general kind of job, not a specific job (AERA/APA/NCME, 1985, p. 64). [It is important to point out that teachers, unlike doctors or lawyers do not receive a *totally* generic license to practice. They are licensed for particular grade ranges, subject fields or specialties. Clearly the domain of knowledge should be specific to the *license* applied for (Madaus & Pullin, 1987). As mentioned earlier, whether there exists a set of pedagogy questions that are necessary across all grade levels is an open question.] (2) The domain does not need to cover the total set of tasks that are performed on the job — only those tasks where, if done incorrectly, some harm may come to the public. (3) The domain should balance criticality and frequency; it need not have the subcategory weights proportionate to the amount of time one spends on that subcategory on the job.

If the job-analysis survey obtains responses to both the criticality and frequency of the objective (task), one has to determine how to weight the two pieces of information. Lahey (1985) and Schmeiser (1987, 1988) have found that different methods of obtaining composite scores produced a very high degree of similarity of weights. However, Kane et al., (1986) did find differences across the algorithms used in their study. As all researchers state — more research is needed.

An area that, to the authors' knowledge, has remained unstudied is whether differences in domain specifications have much impact on the final scores. If indeed the resulting test is measuring a unidimensional domain, it may make little difference if one subarea receives more weight than another. However, if the test is multidimensional, differential weighting would likely have greater impact.

The best advice the authors can give at the present time is: Read the research literature (see the references above); recognize that for licensure, criticality should be considered more important than frequency; and then make the best judgment possible about the subdomain weights. Be prepared to defend the judgment in court!

Item Writing and Review

The writing of multiple-choice items is basically no different for the purposes of teacher-licensure tests than for any other test given to educated adults. Some experts advocate that item writers follow guidelines commonly called item specifications (Popham, 1984), while others prefer to write items directly from the task statements. Either way, it is essential that the written items be reviewed by a panel (or panels) of experts. The items should be reviewed against guidelines like those found in any test and measurement book.

Specific review procedures vary, but all would likely request that the reviewers to make judgments about the accuracy of the item, whether it is congruent with the task (objective, competency) it is intended to measure, whether it is significant, and whether it is lacking in bias. One could have the same committee of experts do this, or one could have different experts review the items for different characteristics. It would be preferable for there to be separate panels dealing with content/accuracy, technical qualities of the items, and potential bias. The qualifications of the three panels should obviously differ.

Here again the wording to the panels needs careful consideration and more research is needed. Madaus and Pullin (1987) offer a series of what-if questions related to the item review and content validity stages of test development.

Field Test

It is both wise, and a common practice, to try out items before they are actually used in an examination. Item try-outs have the potential for informing the test builders about the quality of the items. Several terms are used which suggest different degrees of trying out the items. These terms include *small-scale pilot* and *field test*.

In general, the term *small-scale pilot* suggests that the items will be tried out on 5-to-30 individuals and that the items may not actually be placed together into a draft total test. An advantage of employing this first step is that it is a relatively inexpensive way to spot potentially faulty items. The small-scale pilot generally involves interviewing prospective examinees to determine whether they understood the item, if not why not, or why they chose certain distractors. The reactions of a small number of potential examinees can shed valuable light on possible weaknesses in some items.

A limitation of any small-scale pilot is that one cannot compute meaningful item statistics from the data obtained. Thus, many advocate doing a field test where a sample test is actually produced and administered under conditions that approximate the proposed actual test administration conditions. With a large enough sample, one can compute various item and test statistics as well as conduct bias studies.

While theoretically field tests can be useful, there are several factors that mitigate against the usefulness of field tests for licensure examinations. There are three major problems: (1) sample representativeness (and availability); (2) motivation of the test takers; and (3) preparation for the test.

Considerable research has been done on the accuracy of field-test data. It is not reassuring. Carlson (1987) found that "with few exceptions, field-test performance underestimated real-test performance by a statistically significant and often practically significant amount" (p. xii). Not only are the difficulty indices of the items underestimated, the discrimination indices will be estimated incorrectly and the degree or direction of incorrectness cannot be determined. Thus, there are some very real problems with using field-test data to choose items — or to determine how to revise them.

Nonetheless, there is an ethical issue of initially using a test to make important individual and societal decisions without a knowledge of how individual items behave. A practical solution to the problem is to make the first real administration of the test a "field test." This can be done by including more items in the test than one plans on using and then using the information from the item statistics to choose which items will actually count. If this approach is taken, the developer must be willing and able to drop items that do not pass muster. It is not ethical to argue that items that perform badly (undefined here) need to be retained in order to meet test specifications if the decisions are based on a total score and the test is supposedly unidimensional. While the procedure of combining the field test with the first administration appears to offer the best solution to the field-test problems, it does have drawbacks: while some measurement experts are sanguine with this approach, others remain quite dissatisfied.

Bias Studies

Conducting empirical bias studies is even more difficult than simply conducting field tests. In bias studies, one has not only to have an appropriate set of individuals taking the test who are well motivated and who have prepared as extensively as actual test takers — it is also important to have a sufficient number of individuals representing the relevant protected classes. While bias studies should be done, no one procedure is required. The National Teacher Examination handles bias by analyzing all the items on the test for each administration using the Mantel-Haenszel technique. Items flagged are then reviewed and those judged to be biased are not counted in the scores. (For some subject-matter fields, an insufficient number of test takers at any given administration would negate the usefulness of this approach.)

Reliability Data

Licensure tests are used to make dichotomous decisions, and as the *Standards* state, estimates of the consistency of the decisions are needed (AERA/APA/NCME, 1985, p. 20). Further, it is useful to have an estimate of the standard error of measurement at the cut score. While the *Standards* do not specify any particular formula, measurement professionals would likely view the Subkoviak, Huynh, and Marshall procedures as acceptable (Subkoviak, 1984).

The *Standards* indicate a clear preference of indices based on a threshold-loss function. Hambleton and Novick (1973) and Traub (1986) also argue for the threshold-loss function for licensure tests. Mehrens (in press), while not taking a position, states the possible case for using a squared error loss function. Berk (1984), Subkoviak (1984) and Brennan (1984) discuss this whole area in much greater detail.

It is also considered useful to estimate the reliability of the domain score estimates — consistency across parallel or randomly parallel test forms. As Traub (1986) points out, although such an estimate is not required by the *Standards* for licensure tests, it does provide useful information. It does not replace the need for one of the estimates discussed in the previous paragraph.

Validity Data

Validity of licensure exams is a particularly thorny topic. There is a good deal of disagreement within the measurement community about validity of licensure tests; in fact, the coauthors of this chapter do not even agree on many of these validity issues. Limited space does not allow each of them to articulate their views as fully as they would like. Thus, they urge the reader to read their previous writings to obtain a more thorough understanding of their disagreement concerning the validity of teacher-licensure tests (Haney, Madaus, & Kreitzer, 1988; Madaus, (in press); Madaus & Pullin, 1987; Mehrens, 1986, 1987a, 1987b).

Disagreements aside for the moment, the authors begin with a few fundamental validity concepts. By way of definitions, the *Standards* state that validity:

... refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences (AERA/APA/NCME, 1985, p. 9).

Validity is a unitary concept, but validity evidence can be accumulated in many ways. It is useful to categorize validity evidence as content-related, criterion-related, and construct-related although as the *Standards* remind us, "rigorous distinctions between the categories are not possible" (AERA/APA/NCME, 1985, p. 9). The *Standards* provide appropriate general descriptions of the three

types of validity evidence and of the type of evidence that supports (or refutes) the validity of certain inferences. The need for validity evidence for teacher-licensure tests is no different from the need for such evidence for any other kind of data used to make licensure decisions.

Because evaluation is an ongoing process, accumulating evidence over time, anyone interested in the characteristics of a successful teacher is necessarily interested in the content, criterion-related and construct-validity evidence of any measures of those characteristics. The authors disagree with respect to the amount and quality of validity evidence it is reasonable to expect prior to using a test for licensure purposes and the responsibility of the licensure board (as opposed to researchers) in conducting ongoing validity studies.

Content Validation

Everyone agrees a licensure test should have some content-validity evidence. Content-validation evidence is established during the test-construction process (Cronbach, 1980). The quality of validation evidence is determined by examining the appropriateness of the test-construction procedures. If the original list of competencies was developed by experts, if the job-analysis survey was conducted adequately, if the test specifications were developed from the results of the first two steps, and if the items were written and validated in a satisfactory manner, then it is a reasonable inference that the test is measuring the appropriate content. If the above named steps were not followed or were not executed improperly, then the evidence is questionable. Reasonable professionals may have different standards regarding the adequacy and quality of content validity evidence.

Criterion-Related Validity Evidence

As the authors have mentioned, licensure tests should be designed to protect the public. Based on scores from these tests, an inference is made that a person who falls below the cut score lacks of minimum knowledge or skills, as Mehrens believes, to ensure the public welfare or, at Madaus states, to be minimally successful in the classroom. No inference is made that one who scores *above* the cut score *will* perform well or be effective (see Medley's definitions).

Madaus feels strongly that one should gather criterion-related evidence to determine whether those who score below the cut score can be "minimally successful" in the classroom. If so, there is evidence *disconfirming* the validity of the cut score for the test and perhaps disconfirming the validity of the test itself. (See Madaus, in press, for some specific suggestions on how to conduct such a study.)

Mehrens would agree at a *theoretical* level that evidence showing individuals who score below the cut score are "minimally successful" or do "adequately protect the probable welfare" would be disconfirming. However, he raises objec-

tions on practical and ethical grounds. He agrees with Kane that "attempts to collect such data might be considered unethical (and probably illegal) in many professions" (1984, p. 5). Further, Mehrens discusses the criterion program at great lengths and suggests that:

Even if such data were gathered, a lack of a relationship could well be due to our inability to detect those practitioners who are incompetent and causing harm to the public (Mehrens, 1987b, p. 212).

... harm has been inflicted if the student learns less than the optimal amount due to a teacher's lack of knowledge about basic communication skills, the subject matter taught, or the pedagogy of teaching. To measure that harm in a research study would indeed be difficult. (Mehrens, 1987b, p. 224).

Mehrens concluded that if a study finds that "a lack of necessary knowledge does not harm pupils the public may well doubt the results, and so might I" (Mehrens, 1987b, p. 224).

Madaus and Mehrens would agree that practitioners could be incompetent and causing harm because of characteristics not related to the test. However, the primary issue for Madaus is that the inference being made is that teachers who fail the test cannot be successful in the classroom. He feels this is an empirical question that criterion-related validity studies can answer.

Construct Validation

It is imperative in making an inference from a test score that one should know *what* the test measures. Evidence can and should be gathered to inform users regarding what is being measured. As discussed earlier, teacher licensure tests measure competence. Some of the disagreements about construct-validity evidence for teacher-licensure tests hinge upon whether or not teacher competence is indeed a construct. (For further statements of the authors' differing views about both the need for construct validity and the quality of construct validity evidence see Mehrens, 1987a, 1987b; and Madaus & Pullin, 1987.)

The *Standards* suggest that when one is measuring a construct, one is measuring a *psychological* characteristic. It is called a *construct* because it is a "theoretical construction about the nature of human behavior" (AERA/APA/NCME, 1985, p. 9). Mehrens, following Medley's notion that competencies refer to *specific things* that teachers *know, do, or believe*, submits that competencies are not *theoretical* constructions about the *nature* of human behavior. Note also that the *Standards* (AERA/APA/NCME, 1985, p. 63) suggest that licensure tests are to be "limited to the knowledge and skills necessary to protect the public" (p. 64). *Abilities* is left out of the statement.

Madaus & Pullin (1987) and Linn (1980) argue that competence is a construct. In a more general vein, to cite Messick, "Statements regarding inability or

incompetence require inferences, and the very use of the term inability invokes constructs of attributes and process, whereas a content-valid interpretation would be limited to the outcomes" (Messick, 1975, p. 960). Mehrens & Lehmann (1987) suggest that the Madaus and Linn positions are subtle ones that may at times be useful and at times counterproductive:

The difference between performing in a certain way [on a test] and having the ability (a construct) to perform in a certain way is not one we typically make in everyday language. To infer that a level of performance represents some *hypothetical* construct may only encourage mysticism (p. 82).

Reporting Results

The authors consider two basic questions in this section: (1) Who are the score recipients?, and (2) What should they be told? Madaus and Mehrens agree with the general position taken by Vorwerk and Gorth (1986) that the examination results should be reported to four parties: individual examinees; the colleges and universities the certification applicants attended; the state, which must determine whether certification should be granted; and finally, the public (who should receive aggregate results). The authors submit that the colleges and universities should also not have scores that are identified at the individual level.

The scores reported at the individual level (to the applicants and the state) should only be reported as *pass* for those who score above the cut score. If the actual score were reported, it could lead to inappropriate ranking (Shannon, 1986, p. 36). Remember that a licensure examination is not an employment examination, and inferences about degree of success for those who score above the cut score have not been validated. Thus, there is no evidence that would justify an employer selecting among applicants based on how much they score above the cut score. If the score is reported either to the individual or to the state, it could be used for that purpose.

Scores reported in the aggregate to universities or the public could be based on the actual (continuous) scores. It might well be useful for policy makers to know the total distribution of scores, not just the percent of individuals who scored above the minimum cut score.

There is more debate with respect to the issue of reporting subscores. In general, licensure tests are not primarily designed to be diagnostic. They are designed to categorize individuals into two groups and should have been designed to maximize the reliability of the decision at the cut score. However, tests are built from content outlines that permit the breakdown of the scores into subtest scores. Many argue that reporting the subtest scores would be useful both as a guide for those who have not passed and wish to retake as well as for those who have training/education responsibilities for potential applicants.

Summary and Conclusions

In this chapter, the authors discussed briefly the purpose of licensure, the advantages and limitations of conventional multiple-choice tests, and what inferences are to be drawn from the test scores. Next they discussed, in some detail, test-development procedures that should be employed. Finally, they covered the value of field testing, the importance of bias analyses, obtaining evidence regarding reliability and validity, and the reporting of results.

While the careful reader will note that the authors do not totally agree on every detail discussed, the areas of agreement far exceed the areas of disagreement.

References

- AERA/APA/NCME. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Psychological Association. (1980). *Principles for the validation and use of personnel selection procedures*. Washington, DC: Author.
- Berk, R. A. (1984). Conducting the item analysis. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. (pp. 97-143). Baltimore: The John Hopkins University Press.
- Bloom, B. S., Madaus, G. F., & Hastings, T. J. (1981). *Evaluation to improve learning*. New York: McGraw-Hill.
- Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. (pp. 292-334). Baltimore: The Johns Hopkins University Press.
- Burns, R. L. (1985). Guidelines for developing and using licensure tests. In J. C. Fortune & Associates (Eds.), *Understanding testing in occupational licensing*. (pp. 15-44). San Francisco: Jossey Bass.
- Carlson, R. E., Jr. (1987). *The accuracy of field-test data in predicting future examinee performance on student and teacher competency tests*. Unpublished dissertation, Los Angeles, University of California.
- Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade. New direction for testing and measurement*. (pp. 99-108). #5. San Francisco: Jossey Bass.
- Elliot, S. M., & Nelson, J. (1984). *Blueprinting teacher licensing tests: Developing domain specifications from job analysis results*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Equal Employment Opportunity Commission. Civil Service Commission, Department of Labor and Department of Justice. (1978, Aug. 25) Uniform guidelines on employee selection procedures. *Federal Register*, 43, (166). 38290-38315.
- Guyton, E. & Farokhi, E. (1987). Relationship among academic performance, basic skills, subject matter knowledge, and teaching skills of teacher education graduates. *Journal of Teacher Education*, (Sept./Oct.) 37-42.
- Hambleton, R. K. & Novick, M. R. (1973). Toward an integration of theory and criterion-referenced tests. *Journal of Educational Measurement*, 10, 159-170.
- Haney, W., Madaus, G., & Kreitzer, A. (1988). Charms talismanic: testing teachers for the improvement of American education. In E. Rothkopf (Ed.), *Review of Research in Education*, 14, 169-238. Washington, DC: American Educational Research Association.
- Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29, 340-362.
- Kane, M. T. (1984). *Strategies in validating licensure examinations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. (1986). *A study of nursing practice and role delineation and job analysis of entry-level performance of registered nurses*. Chicago, IL: National Council of State Boards of Nursing, Inc.
- Lahey, M. A. (1985). Personal communication.
- Lavelly, C. D. (1987, August). *Phase I: Competency, skill, and item specification development*. Initial certification and career achievement program. Final report. Tampa, FL: University of South Florida.
- Linn, R. L. (1980). Issues of validity for criterion-referenced measures. *Applied Psychological Measurement*, 14, 547-561.
- Madaus, G. F. (in press). Legal and professional issues in teacher certification testing: A psychometric snark hunt. In J. V. Mitchell, B. Plake, & S. Wise (Eds.), *The assessment of teaching* (pp. 209-259). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Madaus, G. F. & Pullin, D. (1987). Teacher certification tests: Do they really measure what we need to know? *Phi Delta Kappan*, 69(1), 31-37.
- Medley, D. M. (1982). *Teacher competency testing and the teacher educator*. Charlottesville: Association of Teacher Educators and the Bureau of Educational Research, University of Virginia.
- Mehrens, W. A. (1986). *Validity issues in teacher competency tests*. Gainesville: University of Florida. Institute for Student Assessment and Evaluation.
- Mehrens, W. A. (1987a). *Issues in teacher competency tests*. Prepared for the Commission on Testing and Public Policy. Graduate School of Education, University of California, Berkeley.
- Mehrens, W. A. (1987b). Validity issues in teacher competency tests. *Journal of Personnel Evaluation in Education*, 1, 195-229.
- Mehrens, W. A. (in press). Assessing the quality of teacher assessment tests. In B. S. Plake, S. L. Wise, & J. V. Mitchell (Eds.), *Teacher assessment*.
- Mehrens, W. A. & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology*, (3rd ed.). New York: Holt, Rinehart & Winston.
- Mehrens, W. A. & Lehmann, I. J. (1987). *Using standardized tests in education*, (4th Ed.). New York: Longman.
- Messick, S. A. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-966.
- Nitko, A. J. (1983). *Educational tests and measurement*. New York: Harcourt, Brace & Jovanovich.
- Pechione, R. L., Tomala, G., & Forgione, P. D., Jr. (1986). Building a competency test for prospective teachers. In W. P. Gorth and M. L. Chernoff (Eds.), *Testing for teacher certification*. (pp. 99-113). Hillsdale, NJ: Lawrence Erlbaum.
- Popham, W. J. (1984). Specifying the domain of content or behaviors. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. (pp. 29-48). Baltimore: The Johns Hopkins University Press.
- Porter, A. (1988, April). *Understanding teaching: A model for assessment*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.
- Pottinger, P. S. (1979). Competence testing as a basis for licensing: problems and prospects. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education* (pp. 28-46). Washington, DC: National Council on Measurement in Education.
- Pratt, D. (1977). Predicting teacher survival. *The Journal of Educational Research*, 71(1), 12-18.
- Rosenfeld, M., Thornton, R. F., & Skurnik, L. S. (March, 1986). *Analysis of the professional function of teachers: Relationships between job functions and the NTE Core Battery*. Research Report 86-8. Princeton, NJ: Educational Testing Service.
- Rudner, L. M. (1988). Teacher testing: An update. *Educational Measurement: Issues and Practice*, 7(1), 16-19.
- Schmeiser, C. (1987, April). *Effects of translating task analysis data into test specifications*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.
- Schmeiser, C. (1988, April). *Effects of statistical and judgmental weightings of task analysis data on test specifications*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Schmeiser, C. & Estes, C. (1986, April). *Translating task analyses results into test specifications: Clarity or confusion?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

- Shannon, G. A. (1986, April). *Usefulness of score interpretive information for examinees who fail criterion-referenced tests*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Shimberg, B. (1982). *Occupational licensing: A public perspective*. Princeton, NJ: Educational Testing Service.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14.
- Shulman, L. S. (n.d.). *Knowledge and teaching: Foundations of the new reform*. Paper prepared for the Task Force of Teaching As a Profession. Carnegie Forum on Education and the Economy, New York.
- Subkoviak, M. J. (1984). Estimating the reliability of mastery-monmastery classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. (pp. 267-291). Baltimore: The Johns Hopkins University Press.
- Traub, R. E. (1986). Review of test reliability assessment procedures and results. In R. M. Jaeger and J. C. Busch (Eds.), *An evaluation of the Georgia Teacher Certification Testing Program* (chapter 5). Greensboro, NC: Center for Educational Research and Evaluation. University of North Carolina.
- United States Department of Health, Education and Welfare. (1971). *Report on licensure and related health personnel credentialing*. PHEW publication 72-11. Washington DC: Author.
- Vorwerk, K. E. & Gorth, W. P. (1986). Common themes in teacher certification testing program development and implementation. In W. P. Gorth & M. L. Chernoff (Eds.), *Testing for teacher certification* (pp. 35-43). Hillsdale, NJ: Lawrence Erlbaum.
- Yalow, E. S. & Collins, L. J. (1985). *Meeting the challenge of content validity*. Presented as part of a symposium session "The assessment boomerang returns: Competency tests for Educators" at the annual meeting of the American Educational Research Association, Chicago, IL.
- Zieky, (1988, April). *Performance tests for teacher certification: Panacea or Pandora's box?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

17. Performance Tests, Simulations, and Other Methods

Edward H. Haertel

Performance testing and simulation lie somewhere between observing teachers in their classrooms and tallying the circles darkened on their answer sheets. These methods are only now being explored, but the prototypes under development promise to assess a greater range of teaching knowledge and skills than is possible with multiple-choice tests and to do so with greater reliability and efficiency than have been attained in classroom observations of comparable duration.

This chapter discusses teacher assessments conducted outside of actual teaching situations, but with tasks, settings, or materials designed to simulate those of actual practice. Teachers might be asked to critique a textbook, for example, plan a lesson, discuss or correct student homework or papers, comment on a videotape of another teacher's performance, or discuss the use of specialized instructional materials such as historical documents or Cuisenaire rods. Some of these exercises, such as planning a lesson or correcting homework papers, are performance tests, closely simulating tasks teachers are actually required to do. Others, such as analyzing the strengths and weaknesses of a textbook and evaluating its suitability for use with different kinds of learners, may represent tasks performed only rarely if at all by most teachers, but may nonetheless call for types of knowledge and skill central to emergent conceptions of the knowledge base of teaching (Shulman, 1987b). In this chapter, all of these assessments will be referred to as *performance exercises*.

Teacher-performance exercises have been created to measure domains of knowledge and skills that appear critically important for good teaching, but that may be difficult or impossible to measure with conventional, multiple-choice examinations. It is an advantage of these exercises that they look and feel like

AUTHOR'S NOTE: The author gratefully acknowledges the many helpful comments of Stephen Klein and Leon Smith on an earlier draft of this chapter. The chapter has also benefited from the author's many productive discussions with the faculty and students on Stanford University's Teacher Assessment Project, directed by Lee S. Shulman and sponsored by the Carnegie Corporation of New York.

teaching, but face validity has not been the primary consideration of their design. A greater degree of fidelity to the materials and activities of teaching is no more than a means to the end of eliciting a kind of craft knowledge that goes beyond command of subject matter or general pedagogical principles to an understanding of how to impart a particular subject matter to learners. This content-specific pedagogical knowledge (or pedagogical content knowledge) includes an awareness of effective instructional activities for different learners; typical preconceptions students may bring to the learning situation and stumbling blocks they may encounter; effective metaphors and analogies; motivational devices; the strengths and weaknesses of available curriculum materials and the situations in which each may be effective; and the relation of the subject matter to the broader context of the students' entire curriculum, past, present and future.

Among the performance exercises now being developed are one requiring teachers to present a lesson to a small group of students, another in which a teacher examinee views and discusses a videotape of another teacher's performance, and yet another in which a teacher responds to students' scripted questions. These exercises bear some resemblance to the microteaching activities often employed in teacher-education programs (Allen & Ryan, 1969), but differ from microteaching in their purpose and conceptual basis. Performance exercises are tools for assessment, not training. Moreover, the performance exercises requiring a lesson presentation or videotape critique are designed to probe a teacher's overall lesson design and presentation, and call for much more than the execution of specific target skills. Microteaching, at least in its original conception, focused on identifying, shaping, and reinforcing specific teaching behaviors such as set induction, reinforcement, or cue attendance (McDonald, 1973).

The performance exercises considered here are also unrelated to the "performance tests" pioneered at UCLA in the late 1960s and early 1970s (McNeil, Alkin, & Klein, 1974; Popham, 1968, 1971). These were assessments in which examinees were given carefully designed instructional materials to use in teaching actual pupils in school settings. The examinees were given explicit behavioral objectives that their students were to meet, and student achievement tests keyed to those objectives were prepared. Teacher effectiveness was evaluated by looking at posttest scores or pretest-to-posttest gains for the students instructed. The performance exercises described in this chapter differ from those earlier tests in that they are designed for use outside of regular classroom settings and are scored on the basis of what the examinee says and does, rather than the test scores of the students instructed. Most of the performance exercises now being developed do not even involve any actual students. Nonetheless, these earlier performance tests raised issues that are very much relevant to the present exercises, including the problems of defining and quantifying teaching effectiveness, and its generalizability over lesson content, student characteristics, and other dimensions. (See also Shavelson & Dempsey-Atwood, 1976).

Rationale for New Forms of Teacher Assessment

Multiple-choice, paper-and-pencil teacher-licensure tests are designed to measure knowledge and skills important to teaching (Educational Testing Service, 1984; Mehrens, 1986; Nelson, 1985; Poggio, Glasnapp, Miller, Tollsfsen, & Burry, 1986; Quirk, Witten, & Wineberg, 1983; Rebell, 1986). Most of these examinations measure basic literacy, numeracy, and communications skills; general knowledge; knowledge of general pedagogical principles; or knowledge pertaining to specific subject areas, groups of learners, or specialized educational services (Haertel, 1988a; Sandefur, 1986). It appears clear that these areas of knowledge and skill are important to teaching, and so paper-and-pencil tests can be justified on the basis of their content validity. It is also clear that these tests leave many important areas of teacher knowledge and skill unmeasured. Paper-and-pencil tests are inherently limited to examining knowledge *about* a topic.

Performance exercises can by no means measure all of the kinds of knowledge and skills teachers require, any more than paper-and-pencil tests can. Moreover, even some things that they can measure may be tested more efficiently in other ways. In addition, as with almost any deliberate examination, performance exercises are tests of maximal performance. Even if they can show what a teacher is *capable* of doing, they cannot reveal what that teacher *typically* does. Despite these limitations, these exercises may be valuable as complements to more traditional types of tests, providing information about important teaching proficiencies that are otherwise difficult or impossible to assess. Performance exercises are being developed in the hope that they can provide better information about teachers' ability to *apply* what they know — *knowing how* as opposed to *knowing that*.

At least three influences have contributed to the recent interest in more valid and comprehensive methods of teacher assessment, including performance-based approaches. First, teacher testing has expanded greatly over the past decade, as policy makers have instituted more extensive testing requirements in an effort to improve the quality of the teaching force (Lehmann & Philips, 1987; Rudner & McKinney, 1987; Sandefur, 1986). As tests have been more widely used to reach important kinds of decisions, their validity has come to matter more and has come under closer scrutiny.

Second, teacher tests are exerting an increasing influence on teacher-education curricula as more and more states require them on licensure (Southern Regional Education Board, 1985; Spencer, 1986; Wright, 1980). If this influence is to be salutary, then the tests should be enough like teaching that if teacher education programs made direct attempts to improve test performance, they would also be imparting knowledge and skills of direct value to teachers in the classroom (Aburto & Haertel, 1986; Shulman, 1987a). Teacher education programs have recently come under strong criticism (e.g., The Holmes Group, 1986). Many such programs might be improved if in addition to learning the academic knowledge covered by present paper-and-pencil tests, students were

given more practice in planning lessons, critiquing curriculum materials, responding to student questions, or other activities required by performance exercises. A comprehensive system of examinations, including performance exercises could not only guide teacher education curricula, but could also help to codify the knowledge base of teaching.

Finally, certification based on a rigorous, comprehensive, and valid system of assessments is one of the hallmarks of a profession, and efforts are underway to establish such a certification process for teachers. The 1986 report of the Carnegie Task Force on Teaching as a Profession, *A Nation Prepared*, called for the creation of a National Board for Professional Teaching Standards "to establish standards for high levels of competence in the teaching profession, to assess the qualifications of those seeking board certification, and to grant certificates to those who meet the standards" (Task Force on Teaching as a Profession, 1986, p. 62). Teacher *certification*, as used by the National Board for Professional Teaching Standards, refers to a formal recognition of mature and excellent practice by members of the teaching profession itself. It is distinct from *licensure*, which is a governmental function designed to assure a minimum level of proficiency necessary for safe and effective practice by entry-level teachers. There is some confusion in terminology, because a state teaching license is generally referred to as a teaching *certificate*.

Through 1988, most of the research and development on teacher-performance exercises has taken place at Stanford University and at the Connecticut State Department of Education. At Stanford University, Professor Lee S. Shulman directs the Teacher Assessment Project (TAP), sponsored by the Carnegie corporation. The TAP is carrying out basic and applied research on teacher assessment, intended in part to assist the National Board with its work. The Connecticut State Department of Education is working to develop and implement the Connecticut Teacher Assessment Program, which will feature performance exercises along with multiple-choice examinations for teacher licensure (Pechione, Baron, Forgiione, & Abeles, in press).¹ Performance exercises for teacher licensure would be similar to those for teacher certification, but, of course, candidates for licensure would be held to less rigorous standards. Professor Gaea Leinhardt and her colleagues at the University of Pittsburgh and the Learning Research and Development Center (LRDC) have been actively involved in assisting both the Stanford and the Connecticut efforts. Related work has also been conducted at the University of California at Berkeley, Michigan State University, and elsewhere.

Performance Exercises in Assessment Centers and Professional Examinations

The performance exercises being developed for teachers are related to those used in assessment centers for the selection and evaluation of administrators or business executives. In an assessment center, a team of trained examiners admin-

isters exercises to groups of candidates over a period typically from one to three days, scores them, and rates each candidate with respect to a set of dimensions judged critical for job performance (Bray & Grant, 1966; Byham, 1986; Moses & Byham, 1977; Thornton & Byham, 1982). These dimensions are conceived as clusters of specific, observable behaviors. Tasks such as in-basket exercises, leaderless group discussions, interviews, management games, fact-finding exercises, or presentations are each rated with respect to some subset of the chosen dimensions, and these ratings are then integrated by the examiners to arrive at overall appraisals of the candidates.

Assessment centers have not yet been developed for teachers, but a number of centers for selecting high school principals have been established under the National Association of Secondary School Principals (NASSP) Assessment Center Program (Hersey, 1986). Principal assessment centers have also been created by state and local education agencies (e.g., Michaels, 1983; Shilling, 1986) and at universities (e.g., Landholm, 1986), some in collaboration with the NASSP program.

Despite the attempt to define performance dimensions as clusters of specific behaviors, many of the exercises used in typical assessment centers, including centers for principal selection, appear to measure broad traits, not specific job knowledge or skills. Indeed, the nature of the dimensions addressed by assessment centers, and the grounds for their claims of content validity, are topics of current debate (Klimoski & Brickner, 1987; Sackett, 1987). An effort is made to match the traits assessed with those likely to be required in the particular position to be filled, but in fact, the exercises used are substantially similar across a wide variety of corporate settings. Even in assessment centers for principals, most of the exercises closely resemble some of those used to select managers and administrators elsewhere. There are several reasons for this generic quality of many assessment center activities. The specific requirements of an executive or administrative position may be ill-defined, or the intent may be to evaluate candidates for a range of possible future positions. There may be a reluctance to penalize applicants lacking some particular kind of work experience, especially when individuals from outside the firm are being considered. Exercises that are not overly specific may be expected to generalize to a broader range of tasks and situations. The generality of assessment exercises may also reflect the relative ease of adopting or adapting existing exercises for new assessment centers, as opposed to creating entirely new exercises.

Many of the teacher-performance exercises now being developed bear some resemblance to typical performance center exercises, but they are designed to elicit specific, expert teaching knowledge and skill. On exercises designed for use in certification of experienced, practicing teachers, it is appropriate to expect examinees to demonstrate substantial expertise. Even the Connecticut exercises, which are intended for use early in teachers' careers, will be administered to examinees who have acquired sufficient supervised practice in the classroom that specific teaching knowledge may be expected. None of the teacher perfor-

mance assessments being developed or contemplated would be suitable for selecting applicants to teacher education programs. They are measures of acquired knowledge and experience in the art and science of teaching.

The performance exercises employed in examinations for other professions also appear to measure specialized knowledge and skills, rather than general dimensions. They often take the form of complex paper-and-pencil exercises (Lareau, 1985; Aburto & Haertel, 1986). The Architectural Registration Examination, for example, includes a graphics component in which candidates must produce a sample design that is evaluated by trained graders. The California Bar Examination includes two three-hour legal problems for which candidates receive a library of background material, including relevant cases, correspondence, and supporting documentation. Based on these materials, they may be asked to prepare memoranda, briefs, or letters, perhaps a memo to a senior partner setting forth a possible theory of a case. The National Board of Medical Examiners has conducted extensive research and development on a computer-based examination system, in which a case evolves in response to actions taken by the physician-candidate. Oral examinations in professional certification tests also tend to be content specific (Lareau, 1985). In California, architectural candidates must answer questions about their own portfolio of work, and licensed clinical social workers must discuss a case and alternative treatment plans.

Examples of Teacher-Performance Exercises

As of 1988, there are as yet no teacher-performance exercises in widespread use for teacher selection, licensure, certification, or other purposes of evaluation. However, roughly 20 prototype exercises have been developed and piloted by Stanford University's Teacher Assessment Project (Shulman, Haertel, & Bird, 1988), and more will be forthcoming. To facilitate both exercise development and research on exercise reliability and validity, TAP prototypes were developed in two fairly narrow areas: the teaching of mathematics at the upper elementary level, especially the teaching of fractions; and social studies or United States history at the eleventh grade level, especially the American revolution and the formation of the new government. These exercises themselves were never intended to be part of the examinations developed by the National Board for Professional Teaching Standards, but serve to illustrate some possible formats for exercises the Board, state education agencies, or others may develop later. In this section, several of the TAP prototype exercises are briefly described.

Instructional Vignettes

In the draft technical report on this exercise, Marks (1988a) introduces it as follows, from the point of view of the candidate (i.e., examinee, who was playing the role of a candidate for certification by the National Board):

I greeted the examiner and we sat down together at a small table. She told me to assume I was teaching an average fifth grade class and that four of my students came to see me after school, one at a time, each one with a particular difficulty involving calculators and fractional equivalences. She said my task in the exercise was to discuss how I might handle each of these four situations. (p. 1)

In the first situation, a student had converted $1/3$ to 0.333333 on his calculator and wanted to know if they were the same. The candidate was asked to explain in some detail what things she or he would want to get across to the student and would have the student to do, and what related ideas the candidate and student might discuss. In the next situation, the student wondered why he needed to learn "long ways" of comparing fractions when it's so easy just to convert them to decimals on the calculator. The student in the third situation had used a calculator to solve a recipe problem for homework and had come up with answers such as 5.33 tomatoes and .17 teaspoons of pepper (there was also an obvious calculation error). The last situation involved a class of 33 people picnicking at tables designed for 6 people apiece; the student had answered 5.5 to questions like "How many tables will be needed?"

The whole exercise took less than 45 minutes. It was audiotaped, and the examiner also completed a protocol with information concerning timing, any irregularities in the administration, and notations concerning the candidate's responses. The examiner followed an explicit script, with some allowance for probing of excessively brief or unclear responses.

The "tutoring" format of the instructional vignettes exercise was intended to elicit extended conversations about particular student questions or misunderstandings, in a realistic context that minimized the need for concern with teacher time limitations or management of an entire classroom. It was originally designed to assess four broad categories of teacher knowledge (Marks, 1988a, pp. 13-14). First, in the area of *Basic Competence*, the exercise required the candidate to "[apply] standards of reasonableness, practicality, and appropriateness in interpreting and applying mathematics" (Marks, 1988a, p. 13). Second, under the rubric of *Command of Subject Matter*, the exercise probed the teacher's understanding of the limitations of calculators, detection of students' computational errors, knowledge of concepts and procedures involving fractions, and awareness of the relevance of fraction concepts and procedures to other areas of mathematics. Under the third category, *Content-Specific Pedagogy*, ten subareas were listed, including the teacher's awareness of the purposes and goals of mathematics instruction, and ability to justify mathematics content and to frame

explanations in language a fourth or fifth grader could probably understand. The final category, *Pedagogical Understanding and Performance*, was only weakly represented in this exercise, but covered recognizing and responding to students as individuals and as active learners. The points under these different areas were closely tied to specific questions asked as part of this exercise, and possible acceptable answers were tentatively specified a priori.

In scoring, a candidate's actual responses were compared with anticipated answers, and ratings were assigned following explicit rules. Ratings were then assigned for each category, following guidelines for weighing the totality of evidence presented. The categories are logically distinct, but it is an empirical question whether their separate ratings convey any more information than a single, global rating. As more is learned about this and other exercises, new rating formats will probably be developed.

This exercise is designed to be scored by experienced elementary teachers specifically trained on the scoring system. The time required for scoring is determined largely by the necessity of reviewing the audiotape, although speech compression might be used to shorten that time somewhat. Scoring cannot be reduced to a perfectly objective or clerical task, but preliminary work indicates that moderately high interrater reliability can be attained.

Eventually, an exercise of this kind might be modified so that examinees responded in writing other than orally, to either written or videotaped stimuli. (A videotape presentation would allow better pacing, limiting teachers to a realistic amount of time in which to formulate answers to supposed student questions.) Elimination of face-to-face interaction would improve standardization and would make the exercise considerably less expensive to administer and score, although validity might be lower without the opportunity for the examiner to probe brief or ambiguous responses. Requiring written responses might penalize less fluent writers, but at the same time could eliminate sources of invalidity that arise with interviews.

Topic Sequencing

In this exercise, the candidate first arranges a set of 17 index cards labeled with different, related mathematical topics in a fashion that reflects their mathematical and logical interrelationships, and explains the arrangement. Next, the candidate sequences 8 of the topics for purposes of instruction, and discusses that sequencing. The third section of the exercise calls on the candidate to discuss the problems and successes students might have with specific topics. Finally, the examiner switches the order of 2 of the 8 topics and asks the candidate to explain whether or not this is an acceptable teaching order. Among the topics listed on the cards were "ratios," "common denominators," "ordering fractions" "adding fractions with like denominators," "mixed fractions" and "cross multiplying" (Reynolds, 1988a).

The format of this exercise was adapted from part of an interview protocol developed by Gaea Leinhardt at the Pittsburgh LRDC. Examinees are expected to give a sound mathematical rationale for some possible orderings of topics within a learning unit, to demonstrate awareness of topics likely to cause difficulties for students, and to identify common student misconceptions. Thus, the exercise required knowledge of subject matter, of curriculum, and of upper-elementary students (Reynolds, 1988a).

Documentary History

In this exercise, the candidate is provided with a collection of (retyped) eyewitness accounts of the Battle of Lexington; secondary sources describing the incident, including an account from a high school textbook; and three paintings of the battle; together with a map of Lexington Green as it was on April 19, 1775. After studying these materials for about a half hour, she was interviewed concerning the documents themselves, how they might be used in the classroom, and what instructional goals might be advanced through such uses (Hyink, 1988).

In studies of expert history and social studies teachers, the use of documentary evidence in the classroom had emerged as an important aspect of pedagogical content knowledge. This exercise probed the candidate's awareness of stumbling blocks high school students encounter with the language and format of such documents, the pedagogical purposes they could serve, and the relative value of different documents. Candidates were also requested to describe how they might organize the students to analyze and use the documents over a period of two or four days. The documents were chosen to support several instructional aims, including the illustration of a historical myth and its evolution. (The three paintings were dated 1775, 1859, and 1886, and showed important, systematic, progressive distortions of the actual event.)

Additional Exercises

Additional exercise prototypes have been developed at Stanford, following comparable procedures. They include exercises in both elementary mathematics and high school history in which the candidate presents a lesson prepared in advance to a group of six students and is interviewed about it (Shulman, Nelson-Barber, Aburto, & Mitchell, 1988); exercises in both areas in which the candidate views a videotape of a lesson in progress and is questioned about it (Reynolds, 1988b; Kerdeman, 1988); and exercises in which the candidate must analyze a textbook and answer questions about its strengths and weaknesses, either orally or in writing (Kelley, 1988b; Wineburg, 1988). In mathematics, additional exercises include having the candidate plan and discuss a lesson on the topic of a specified textbook passage (Tyson, 1988); discussing how to teach the equivalence of fractions using egg cartons, paper plates, a clock, or other common

articles (Kelley, 1988a), or using Cuisenaire Rods and a simple instructional computer program (Marks, 1988b); discussing the pedagogical strengths and weaknesses of three computational algorithms (Stein, Marks, & Wolf, 1988); and describing classroom routines for checking mathematics homework (Hilger & Grover, 1988). Additional high school history exercises include the evaluation of student papers written in response to an essay question (McGraw & Wilson, 1988); a lesson planning exercise; an exercise in which the candidate views a brief instructional videotape and then answers scripted questions from three students who are supposed to have seen the same tape; one in which the candidate must prepare packets of materials for a cooperative group learning activity; and an exercise in which several candidates work together to plan a curriculum unit.

Scoring Performance Exercises

The first step in creating teacher-performance exercises is to develop standardized situations for eliciting, observing, and recording behavior that reflects a teacher's craft knowledge and expertise. The second step, equally important, is to develop procedures for deriving reliable and valid scores from the performances observed. Appropriate scoring methods and scales will depend on the level of detail required, from a pass/fail decision to a rich diagnostic profile.

Scoring for the TAP exercise prototypes represented a compromise among several purposes. Although the scoring schemes were influenced by the National Board's eventual requirement for pass/fail judgments, several scoring dimensions were retained, and performances were rated with respect to each dimension on a scale differentiating several levels of performance. Further research may show these dimensions to be so highly correlated that some or all of them are better combined. The scale used for the TAP exercises is shown in Table 17.1.

It was tentatively agreed for purposes of the TAP that a pattern of consistent AA ratings would clearly be acceptable for certification, and a pattern of consistent B+ ratings would not. A-level performance, or an inconsistent profile across exercises or dimensions within exercises would be considered borderline. Note that the TAP scale was designed to make finer distinctions near the top end of the proficiency continuum than the bottom. Of course, the National Board may arrive at an entirely different set of criteria and standards, and is expected to use performance exercises in conjunction with other certification requirements and other forms of evidence.

The experience of the TAP suggests that scoring methods for teacher performance exercises should be neither too holistic nor too fragmentary. Global judgments of the quality of performance are likely to be inconsistent from one judge to another, and difficult to defend. On the other hand, checklists of very small elements of behavior may be unduly sensitive to the total amount of writing or talking by an examinee and are likely to be inflexible, favoring one approach to an exercise over other, equally valid approaches. An intermediate

Table 17.1 Proficiency Scale Developed for Teacher Assessment Project Exercises

AAA	Distinguished	Meets AA standard and adds originality, depth, or charisma deserving special notice
AA	Commendable	Mature and balanced professional
A	Satisfactory	Often rises above the merely adequate level
B+	Adequate	Appears satisfactory for standard responsibilities of teachers
B	Limited	Satisfactory in many of the teacher's responsibilities, but falls short of B+
C	Questionable	Shortcomings are unlikely to be overcome by further experience alone

course assumes that scorers are seasoned practitioners, and relies on their knowledge and judgment, but provides substantial guidance about the dimensions according to which performances are to be evaluated, the procedures to be followed in scoring, and the level of performance to be judged acceptable.

Discussion

Performance exercises for teachers are still in their infancy. Early results are promising, but many issues remain to be resolved. Linn (1986) points out that new testing technologies must compete against the extraordinary efficiency and the high criterion-related validity coefficients of multiple-choice objective tests. Criterion-related evidence has not been central to the validation of licensure and certification tests (AERA, APA, & NCME, 1985), and there is dissatisfaction with paper-and-pencil tests, but it nonetheless remains to be seen whether performance exercises will attain wide acceptance and use. This section raises some practical and psychometric issues that will have to be resolved if these exercises are to fulfill their promise.

The Knowledge Base of Teaching

Teacher-performance exercises have been based on direct observation of classroom teachers, as well as research on teaching, both in general and in the specific subject areas assessed. In the TAP, expert teachers identified through a careful nomination process were observed and interviewed, and several of them were closely involved throughout the project as paid consultants. Some of the nation's leading researchers in the teaching of mathematics and of history and the social sciences were convened in advisory panels, which also included university faculty in the cognate disciplines. A separate Teacher Advisory Board also critiqued the exercises. The TAP's reliance on these multiple sources of information highlights the difficulty of determining just what is comprised by the knowledge base of teaching. There are as yet no textbooks that set forth the kind of detailed pedagogical content knowledge these exercises are designed to

assess, and the conventional wisdom of experienced teachers may include unwarranted beliefs as well as verities.

The lack of a consensual, codified knowledge base poses some difficulties for teacher examinations of any form. Prospective examinees should be able to prepare for these tests, and there should be some authoritative source for deciding whether the responses they offer are correct. A definitive source would also be useful in establishing the content validity of a battery of performance exercises. All that being said, it may be unrealistic to wait for consensus on a knowledge base before assessments are developed. The creation of teacher-performance exercises, even if they are imperfect, can focus attention on neglected areas of teacher expertise, stimulate research, encourage deliberation, and accelerate the appearance of textbooks and case study materials, as well as the improvement of teacher education curricula.

Alternative Modes of Administration

Many exercises initially designed around structured interviews might be converted to a paper-and-pencil or computer-based administration. The advantages of these formats could include cost savings in administration and scoring; improved standardization of testing conditions; greater reliability and objectivity in scoring; and creation of a complete and permanent record of examinee performance that could be reviewed if ever the scoring were challenged.

It would also be much easier to maintain the security of paper-and-pencil or computer-administered tests. There are an enormous number of teachers who may seek Board certification, for example, and if the National Board's tests were individually administered by trained personnel, then of necessity those teachers would be examined over a period of days or weeks. This would raise the possibility that those tested later might obtain some advance information from those tested earlier. Experience with other large-scale performance assessments (e.g., the United States Foreign Service Examinations) suggests that this problem can be controlled to some extent by rotating exercises, but a simpler solution might be to prepare exercises amenable to group administration, and to offer a given form of the test on only one occasion.

Disadvantages of written versus face-to-face examinations might include greater cost in the development of exercises, some limitation in the range of knowledge and skills that could be assessed, and diminished fidelity to actual teaching situations and activities. It is possible that written or computer-based assessments would place greater reliance on ancillary, "test-taking" skills important for doing well on the assessment, but not in the actual work of teaching. In particular, paper-and-pencil tests are likely to place a premium on a higher level of writing fluency skill than teaching itself requires. On the other hand, interviews may similarly depend on ancillary oral expressive skills. The most prudent course may be to use a mix of oral and written performance assessments.

Alternate Forms for Performance Exercises

An operational testing program will require multiple forms of exercises that are at least roughly parallel. Precise calibration of the difficulties of these alternate forms may not be necessary if appropriate statistical models are used, but there should be some assurance that different stimulus tapes for videotape critiques, or different sets of instructional vignettes, or different sets of student essays to be graded are measuring the same dimensions of knowledge and skill. At the very least, alternate forms of examinations will be required from year to year. If only a limited number of candidates can be assessed at any one time, the need for multiple forms may be even greater for reasons of test security.

In the TAP, prototype exercises were intended to illustrate potential categories of exercises, and were designed so that specific topics, instructional materials, and so forth could be changed to generate alternate forms. It remains to be shown, however, that such alternate forms could in fact be developed for all of the exercises, and it is an empirical question whether such forms would function in the same way.

Pedagogical content knowledge is conceived as highly specific, and so performance exercises to measure it have involved the detailed examination of narrow segments of curriculum content. There is no assurance, however, that detailed knowledge of one topic reliably predicts detailed knowledge of others. Generalizability studies will be required to determine both the correlations obtainable among alternate forms, and the number of curriculum topics that must be sampled to generalize about pedagogical content knowledge across an entire area of teaching specialization.

Validity and Reliability

The major impetus for developing teacher-performance assessments was dissatisfaction with the validity of existing tests. The validity of these new assessments has been of concern wherever they have been developed. Validity inheres not in a test itself, but in a specific use of the test. The TAP has studied exercises that might be useful as part of a certification process for experienced teachers, and the State of Connecticut is designing semistructured interviews to be used as part of their Beginning Educator Support and Training (BEST) program (Pechione, et al., in press). Neither of these programs is yet operational, however, and so data are not available for a thorough critique of the tests' validities.

At this time, the strongest line of validity evidence is logical, or content-related. Performance exercises are *not* based on a "job analysis" of the activities teachers perform, and few if any of these exercises sample the routine tasks that consume most of a teacher's time. However, all of them are designed to measure knowledge, skills, and dispositions arguably important for excellent teaching. Successful performance should require knowing and doing things that teachers

must know and do (Pechione, et al., in press; Shulman, 1987a, 1987b; Shulman, Haertel, et al., 1988). Preliminary analyses of data from the small-scale TAP field test indicate that teachers with more experience outperform beginning teachers, as would be expected if the exercises succeed in assessing the knowledge base of teaching (Haertel, 1988b). Experience is not the same as expertise, of course, but the two are strongly associated (Berliner, 1986).

In developing performance exercises and scoring systems, attention must be given to potential threats to validity. As already noted, for example, an interview format reduces demands on candidates' writing skills, but increases the potential influence of verbal fluency on the score obtained. Other threats to validity are associated with the limited contextualization of performance exercises. Different instructional approaches may be appropriate depending on the abilities, experiences, interests, language backgrounds, or instructional histories of learners; on the instructional resources available; the philosophy of the school; or any number of other contextual factors. A major challenge in developing and scoring these exercises is to allow for the diversity of acceptable responses while maintaining standards of quality.

The fact that different situations call for different instructional approaches certainly does not suggest any simple correspondence between superficial learner characteristics and appropriate teaching methods, and so the performance exercises developed at Stanford and Connecticut offer little information about the specific kinds of classrooms or learners the examinee is to have in mind. These exercises are intended to support inferences to performance in classrooms rich and poor, diverse in many ways. This could conceivably place teachers whose experience is limited to only one or a few teaching settings at some disadvantage, but that question cannot be addressed with the data now available.

Reliability is prerequisite to validity. As a preliminary indication of reliability, the teams developing scoring systems for the TAP often had two raters independently evaluate the same protocol. They generally obtained good-to-excellent agreement, but that was after the separate raters had discussed the exercises at length, and collaborated in developing scoring rules. A more stringent test of reliability would depend on scoring by persons less intimately involved in exercise and scoring development.

Reliability will be a function of the number of scorers used, the extensiveness of their training, and the level of detail and specificity in the scoring guides. The reliability of performance exercise batteries will also depend on the number and kinds of exercises included, and the range of instructional content and teaching contexts to which results are generalized. The level of reliability considered acceptable for these exercises will depend on how they are incorporated into the decision process. There appears to be no reason to doubt that with reasonable effort and care, a satisfactory level of reliability can be attained.

Both reliability and validity will depend on clarity of task demands. Examinees must be informed of the bases on which their performance will be evaluated and the level of detail expected in their responses. In some of the TAP prototype

exercises, the field test revealed unintended ambiguities. For example, some candidates used the student-questioning exercise to demonstrate their content knowledge, and others used it as an opportunity to demonstrate their ability to establish rapport with students or to initiate a class discussion. Clearer instructions and scoring criteria would have led to more commensurable performances by different candidates.

As teacher tests have been adapted to new purposes, the need for more valid forms of teacher assessment has increased. Performance tests are a means to improving validity by assessing aspects of teacher knowledge largely excluded from existing paper-and-pencil tests. The assessments being developed are grounded in a strong theoretical model of the knowledge base of teaching, and embody a vision of teaching as it might become. It appears likely that performance exercises will assume increasing importance in the teacher tests of the future.

Note

1. Together with California, Connecticut has taken the lead in forming the New Interstate Teacher Assessment and Support Consortium, which is intended among other functions to share information on performance-based teacher assessments and to coordinate similar development efforts among participating states. Further information about the Consortium may be obtained from the Coordinating Office, located at room CERAS-402, School of Education, Stanford University, Stanford, CA 94305.

References

- Aburto, S., & Haertel, E. H. (1986). *Teacher assessment project study group on alternative assessment methods: Executive summary of the assessment technologies conference* (Report No. 5). Stanford, CA: Stanford University Teacher Assessment Project.
- Allen, D. W., & Ryan, K. A. (1969). *Microteaching*. Reading, MA: Addison-Wesley.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Berliner, D. C. (1986). In pursuit of the expert pedagogue. *Educational Researcher*, 15(7), 5-13.
- Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs*, 80(17, Serial No. 625).
- Byham, W. C. (1986). *Use of the assessment center method to evaluate teacher competencies* (Paper prepared for the Teacher Assessment Project, Stanford University). Stanford, CA: Stanford University Teacher Assessment Project.
- Educational Testing Service. (1984). Guidelines for proper use of NTE tests. In Educational Testing Service, *A guide to the NTE Core Battery tests: Communication Skills, General Knowledge, Professional Knowledge* (pp. 143-148). Princeton: Author.
- Haertel, E. H. (1988a). Assessing the teaching function. *Applied Measurement in Education*, 1, 99-107.
- Haertel, E. H. (1988b, April). *Quantifying the wisdom of practice*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.
- Hersey, P. W. (1986). Selecting and developing educational leaders. *School Administrator*, 43(3), 16-17.

- Hilger, G. S., & Grover, B. W. (1988, April). *Lesson structures: Homework correction (mathematics)* (Draft Technical Report No. M8). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Holmes Group, Inc. (1986). *Tomorrow's teachers: A report of the Holmes Group*. East Lansing, MI: Author.
- Hyink, L. (1988, April). *Documentary history* (Draft Technical Report No. H1). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Kelley, L. (1988a, April). *Opportunity box (mathematics)* (Draft Technical Report No. M4). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Kelley, L. (1988b, April). *Textbook analysis (mathematics)* (Draft Technical Report No. M9). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Kerdelman, D. (1988, April). *Videotape critique (history)* (Draft Technical Report No. H6). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, 40, 243-260.
- Landholm, L. J. (1986). Observations of participants: Center helps one's monitoring of strengths, weaknesses. *NASSP Bulletin*, 70(486), 24-25.
- Lareau, A. (1985). *A comparison of professional examinations in six fields: Implications for the teaching profession*. Stanford, CA: Knowledge Growth in Teaching, School of Education, Stanford University.
- Lehmann, I. J., & Phillips, S. E. (1987). A survey of state teacher-competency examination programs. *Educational Measurement: Issues and Practice*, 6(1), 14-18.
- Linn, R. L. (1986). Barriers to new test designs. In Educational Testing Service, *The redesign of testing for the 21st century: Proceedings of the 1985 ETS Invitational Conference* (pp. 69-79). Princeton, NJ: Educational Testing Service.
- Marks, R. (1988a, April). *Instructional vignettes (mathematics)* (Draft Technical Report No. M6). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Marks, R. (1988b, April). *Teaching tools (mathematics)* (Draft Technical Report No. M5). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- McDonald, F. J. (1973). Behavior modification in teacher education. In C. E. Thoresen (Ed.), *Seventy-fourth yearbook of the National Society for the Study of Education, Part 1: Behavior modification in education*. Chicago, IL: National Society for the Study of Education.
- McGraw, L., & Wilson, S. (1988, April). *Evaluation of student papers (history)* (Draft Technical Report No. H3). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- McNeil, J. D., Alkin, M. C., & Klein, S. P. (1974). A new procedure for examining the issue of school wealth and effectiveness of teachers. *Education and Urban Society*, 6, 497-504.
- Mehrens, W. A. (1986). *Validity issues in teacher competency tests*. Gainesville, FL: University of Florida, Institute for Student Assessment and Evaluation.
- Michaels, K. G. (1983). Our assessment center benefits the entire management team. *Executive Educator*, 5(11), 23-25.
- Moses, J. L., & Byham, W. C. (Eds.). (1977). *Applying the assessment center method*. Elmsford, NY: Pergamon.
- Nelson, E. A. (1985). Review of NTE programs. In J. V. Mitchell, Jr. (Ed.), *Ninth Mental Measurements Yearbook* (Vol. 2, pp. 1063-1066). Lincoln, NE: University of Nebraska-Lincoln Press.
- Pecchione, R. L., Baron, J. B., Forgione, P. D., Jr., & Abeles, S. (in press). A comprehensive approach to teacher assessment: Examples from math and science. In A. B. Champagne (Ed.), *This year in school science 1988: Science teaching - Making the system work* (pp. 191-214). Washington, DC: American Association for the Advancement of Science.
- Poggio, J. P., Glasnapp, D. R., Miller, M. D., Tollefson, N., & Burry, J. A. (1986). Strategies for validating teacher certification tests. *Educational Measurement: Issues and Practice*, 5(2), 18-25.
- Popham, W. J. (1968, June). *Performance tests of instructor competence for trade and technical education* (Final Report, Office of Education Project No. 5-004, Contract No. OE-5-85-051). Los Angeles, CA: University of California at Los Angeles. (ERIC Document Reproduction Service No. ED 027 418)
- Popham, W. J. (1971). Performance tests of teaching proficiency: Rationale, development, and validation. *American Educational Research Journal*, 8, 105-117.

- Quirk, T. J., Witten, B. J., & Wineberg, S. F. (1973). Review of studies of the concurrent and predictive validity of the National Teacher Examinations. *Review of Educational Research*, 43, 89-113.
- Rebell, M. A. (1986). Disparate impact of teacher competency testing on minorities: Don't blame the test-takers - or the tests. *Yale Law and Policy Review*, 4, 375-403.
- Reynolds, A. (1988a, April). *Topic sequencing (mathematics)* (Draft Technical Report No. M3). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Reynolds, A. (1988b, April). *Videotape critique (mathematics)* (Draft Technical Report No. M2). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Rudner, L. M., & McKinney, K. C. (1987). State-by-state descriptions. In L. M. Rudner (Ed.), *What's happening in teacher testing: An analysis of state teacher testing practices* (pp. 43-127). Washington, DC: U.S. Government Printing Office. (U.S. Department of Education, Office of Educational Research and Improvement, Report No. OR-87-502)
- Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology*, 40, 13-25.
- Sandefur, J. T. (1986). State assessment trends. *AACTE Briefs*, 7(6), 12-14.
- Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, 46, 553-611.
- Shilling, J. L. (1986). Developing an operational plan: Maryland's initiative for quality leadership. *NASSP Bulletin*, 70(486), 3-6.
- Shulman, L. S. (1987a). Assessment for teaching: An initiative for the profession. *Phi Delta Kappan*, 69, 38-44.
- Shulman, L. S. (1987b). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57, 1-22.
- Shulman, L. S., Haertel, E. H., & Bird, T. (1988, April). *Toward alternative assessments of teaching: A report of work in progress*. Stanford, CA: Teacher Assessment Project, Stanford University.
- Shulman, L. S., Nelson-Barber, S., Aburto, S., & Mitchell, J. (1988, April). *Teaching a familiar lesson (mathematics & history)* (Draft Technical Report No. M10 & H8). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Southern Region Education board. (1985). *Strengthening minority students and test-taking skills*. Atlanta, GA: Author.
- Spencer, T. L. (1986). Teacher education at Grambling State University: A move toward excellence. *Journal of Negro Education*, 55, 293-303.
- Stein, M. K., Marks, R., & Wolf, K. (1988, April). *Shortcuts and other algorithms (mathematics)* (Draft Technical Report No. M7). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Task Force on Teaching as a Profession. (1986, May). *A nation prepared: Teachers for the 21st century*. Hyattsville, MD: Carnegie Forum on Education and the Economy.
- Thornton, G. C., III, & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York, NY: Academic Press.
- Tyson, P. (1988, April). *Lesson planning (mathematics)* (Draft Technical Report No. H1). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University. (Draft Technical Report No. M1). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Wineberg, S. S. (1988, May). *Textbook analysis (history)* (Draft Technical Report No. H7). Stanford, CA: Teacher Assessment Project, School of Education, Stanford University.
- Wright, S. J. (1980). The survival of Black public school teachers: A challenge for Black colleges and universities. In E. P. Witty (Ed.), *Proceedings of the National Invitational Conference on Problems, Issues, Plans, and Strategies Related to the Preparation and Survival of Black Public Schools Teachers*. Norfolk, VA: Norfolk State University, School of Education. (ERIC Document Reproduction Service Nos. ED 212 565, ED 212 568).

18. Setting Standards on Teacher Certification Tests

Richard M. Jaeger

What's in This Chapter and What Isn't

Test standards are also called *cut scores* or *passing scores*. By whatever label, a test standard defines the lowest score that can be earned by an examinee who passes a test; examinees who earn scores at or above the test standard are said to have passed the test, and examinees who earn scores below the test standard are said to have failed the test.

This chapter is a practical guide to setting standards on certification tests used with practicing teachers or with persons who desire to become teachers. In most states, different tests are used to screen applicants to teacher-education programs and to screen candidates for certification to teach (Rudner, 1987). For simplicity, all such tests will be called *teacher certification tests*. This chapter is *not* a theoretical treatise on standard setting, nor does it provide an exhaustive review of the psychometric literature on standard setting. Several comprehensive reviews of standard-setting methods and some articles on the methodology of standard setting are referenced and described in the final section of this chapter.

As a practical guide to standard setting, this chapter contains a description of several methods that have been used by one or more states to establish standards on teacher-certification tests or to establish standards on tests used to select applicants to teacher-education programs. It includes discussion of such issues as the qualifications of persons who should recommend test standards (we call such persons "judges"), the qualifications of persons who should set test standards (those who recommend standards and those who set standards are often different persons; the latter group usually has statutory responsibility for determining education policy), how to select judges, how many judges to select, how to train judges, what judges should be asked to do, the types of judgment data to collect,

AUTHOR'S NOTE: I am grateful to John Christian Busch, Lorrie Shepard, W. James Popham, and Susan Thomas, for a number of helpful suggestions that improved this chapter materially. I am, of course, solely responsible for any remaining inaccuracies or shortcomings.

how to analyze the judgment data, and how to report and document what you've done.

Serious debates have been held among educators, educational researchers, and measurement experts on the merit, equity, and utility of teacher-certification tests (cf., Folks, 1985; Gifford, 1987; Shepard, Kreitzer, & Graue, 1987). Measurement experts have also debated the appropriateness of setting standards on such tests. These issues are ignored in this chapter, although you will find them discussed in other chapters of this volume. In this chapter, it is assumed that the reader has decided to use a teacher-certification test for some purpose that requires establishing a passing score.

When Test Standards Are Needed

Certification tests and associated passing scores have been used by state governments for such purposes as determining (1) which applicants will be admitted to teacher education programs, (2) which graduates of teacher-education programs will be granted initial certification to teach, and (3) which practicing teachers will be granted certification that allows them to continue teaching. In all of these applications, performance on a test has been used as one criterion that determines whether an examinee will be admitted, initially certified, or granted continued certification to teach.

According to the 1985 *APA Standards for Educational and Psychological Testing*:

The primary purpose of licensure or certification is to protect the public. Licensing requirements are imposed to ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively (p. 63).

Although state education agencies have traditionally issued "teaching certificates," such certificates have played the role of licenses, as without a valid certificate, one cannot legally teach in the public schools of the issuing state. In the argot of industrial and organizational psychology, a certificate is usually awarded by a professional organization, is sought voluntarily by a candidate, denotes satisfaction of professional standards of knowledge or performance, and has no legal status. In contrast, a license is awarded by a government agency or body, denotes satisfaction of minimum standards, and is legally required to practice a profession within the domain of the government agency or body. Therefore, what various state education agencies have labeled "teacher certification tests" actually constitute licensure tests, because typically, an examinee must earn a passing score as a prerequisite to certification, and therefore, to legal employment as a teacher.

Consistent with the 1985 *Standards* definition of licensure tests, most teacher certification tests are designed to assess some aspect of examinees' minimum qualifications to teach. They are not designed to measure teaching proficiency or typical teaching performance. Even when their restricted purpose is acknowledged, it is generally recognized that what teacher-certification tests measure varies by degree among examinees. Therefore, a test standard does not define a clear point of division between those who have what the test measures and those who do not (Glass, 1978; Jaeger, 1982; Linn, Madaus, & Pedulla, 1982; Shepard, 1979, 1980, 1984). Rather, a standard on a teacher-certification test represents the lowest level of test performance that judges define as sufficient to warrant admission, certification, or licensure. This performance level is determined judgmentally and, as some have noted, arbitrarily. However, a test standard need not be defined capriciously. The guidelines provided in this chapter are designed to help the reader avoid capriciousness.¹

What's Involved in Setting a Standard?

Setting a standard on a teacher-certification test merely requires that one or more persons decide on the lowest test score that is consistent with granting admission, certification, or some other positive outcome on a decision in question. In theory, a single person could make that decision capriciously, with little attention to the content of the certification test or the likely performance of examinees. But such a procedure would not be consistent with the specifications of the 1985 *Standards for Educational and Psychological Testing* or with the provisions of the federal *Uniform Guidelines to Employee Selection Procedures* (U.S. Department of Labor, 1978). These documents require that a procedure for setting a test standard be scientifically based and replicable and that the procedure result in a standard that is reasonable, valid, and reliable.

To achieve these goals, a sufficiently large group of well-qualified and well-trained judges should be asked to recommend a test standard on the basis of a procedure that is consistent with the capabilities of the judges and is fully documented. Details of such procedures are discussed in the balance of this chapter.

Who Should Recommend Standards?

Knowledgeable professionals should recommend test standards, and responsible policy makers should set them. What constitutes a knowledgeable professional depends on the purpose for which a test standard is to be recommended and the method that is used to formulate a recommendation. We consider three purposes of testing, in turn.

Determining which applicants will be admitted to teacher-education programs. When teacher-certification tests are used to control admission to teacher-education programs, the intent of policy makers is to admit only those candidates who have sufficient prerequisite knowledge. For example, in the state of North Carolina, to be admitted to a teacher-education program, a candidate must earn passing scores on the General Knowledge Test and the Communication Skills Test of the National Teacher Examinations.

The types of professionals who should recommend standards for this purpose must be familiar with the curricula of teacher-education programs and the demands of those curricula on prerequisite knowledge. Two categories of judges come immediately to mind: faculty members in colleges or universities who provide instruction in teacher-education programs and classroom teachers who are relatively recent graduates of the teacher-education programs. If standard-setting procedures that require judgments about test items are used, judges in these categories should have requisite knowledge about the content domain and, with appropriate training, can be expected to apply that knowledge in a thoughtful and reasoned manner. If standard-setting procedures that require judgments about individual examinees are used, it is necessary to collect judges from the above-named groups who are familiar with the knowledge or performance levels of specific examinees. For example, if a basic writing test is used to screen applicants for a teacher-education program, then judges must know the curriculum of the teacher-education program and the writing proficiencies of a sample of individual applicants well enough that they can accurately predict whether each applicant would perform successfully various writing-dependent tasks expected of students as a part of the teacher-education program.

Determining which teacher education program graduates will be granted initial certification to teach. The intent of policy makers who use certification tests to control initial teacher certification is generally consistent with the 1985 *Test Standards* definition of licensure and certification tests cited earlier. Policy makers want to protect the public from prospective teachers whose knowledge and skills are not sufficient to permit them to teach safely and effectively.

The types of professionals who should recommend standards for this purpose should be familiar with the prerequisite knowledge requirements of the job of beginning teacher in the domains assessed by the teacher-certification tests for which standards are sought. Categories of persons most likely to have such familiarity include teacher educators who supervise student teachers and fully certified practicing teachers who, by virtue of the recency of their initial year of teaching or their supervisory work with beginning teachers or student teachers, are likely to have realistic perceptions of the knowledge and skills required of beginning teachers. Again, if standard-setting procedures that require judgments about individual examinees are used, the judges selected must be familiar with the knowledge or performance levels of specific examinees. For this reason, it might be impractical to use practicing teachers as judges with such standard-set-

ting methods, with the possible exception of those who supervise student teachers.

Determining which practicing teachers will be granted certification to continue teaching. It could be argued that knowledge and skill levels sufficient to protect members of the public from ineffective or unsafe beginning teachers should also be applied to practicing teachers, in which case the same tests and the same test standards might be used to screen candidates for beginning and continuing teacher certification. Conversely, policy makers might want independent consideration of tests and test standards when candidates for continuing-teacher certification are screened. In the latter case, persons selected to recommend standards on tests used for making decisions on continuing certification of practicing teachers should be experts on the prerequisite knowledge and skill requirements of practicing classroom teachers in the domains assessed by the certification test. Practicing teachers who were fully certified and labeled *successful* by their supervisors would be most likely to have trustworthy knowledge of the type required. Teacher educators might also be used as judges for this standard-setting purpose, but their secondary knowledge of the prerequisite skill and knowledge demands placed on experienced classroom teachers might limit the validity of their judgments.

Who Should Set Standards?

Because standards on teacher-certification tests are an important component of education policy and are usually applied uniformly throughout a state, they should be set by persons with statutory authority and responsibility for establishing education policy, such as members of state boards of education. The advice of professionals who have applied a reasonable and well-documented standard-setting procedure in formulating their recommendations should be presented to members of such policy-making bodies.

How Should Judges be Selected?

Populations of judges that could appropriately recommend standards for three applications of teacher-certification testing were described in the previous section. Procedures for selecting judges from these populations are discussed here. Two disparate approaches to selection with somewhat distinct objectives are worthy of consideration. First, it might be possible to specify criteria for identifying the best qualified persons among a population of appropriate judges. If so, selection of judges could include only the best qualified, without regard for differences between the test standards that would be recommended by such a group, and the test standards that would be recommended by a group chosen to represent all appropriate judges. Second, it might be possible to select a repre-

sentative sample from the population of appropriate judges, and thus ensure that the recommendations of sampled judges would reflect the recommendations of the entire population, in an unbiased way. The theoretical question is interesting — Do those “best qualified” to make judgments deserve a greater voice in setting standards than those who are merely “appropriate,” and therefore qualified by some minimum criteria? In practice, selecting the best qualified judges is fraught with difficulties, some of which appear to be insurmountable. Who is to decide which potential judges are best qualified? And who ordained the judges of judges? Are they best qualified to judge the judges? The problem is one of infinite regress. The second choice, attempting to select a representative sample, is also difficult to achieve. Practical constraints often preclude use of the kinds of probability sampling procedures that yield representative samples. However, this situation is not impossible, because workable approximation techniques can often provide useful and defensible results.

The most difficult practical problem in selecting a representative sample is assembling a list of all members of the desired population of judges. For example, if judgments were desired from the state-wide population of fully certified practicing teachers who are no more than a few years beyond their initial year of teaching, a list of such teachers might not be available. In addition, this population of teachers would be geographically dispersed, so it might be impractical to assemble a representative sample of the population to obtain their judgments. To resolve this dilemma, teachers could be selected by using a procedure that preserved some elements of representativeness while providing a feasible solution. Consider as an example, the solution that was used to recommend standards for the Core Battery of the National Teacher Examinations in North Carolina a few years ago (Jaeger & Busch, 1984).

North Carolina contains three identifiable geographic areas that differ markedly in their economics and their sociocultural and racial composition. Representing all three areas proportionately was an important goal in sampling practicing teachers who would serve as judges. The counties comprising each of the three areas were examined carefully. The single county that most closely approximated its area on such variables as median per capita income, median education level of the adult population, proportion of white collar employees, etc. was determined using data assembled by the U.S. Bureau of the Census. Listings of all school districts within 35 miles of the center of the county selected from each area were compiled. The North Carolina Department of Public Instruction provided a list of all fully certified classroom teachers employed by these school districts, who had at least two and no more than five years of teaching experience. A random sample of teachers was selected from each of the lists, so that the size of the sample selected from each area was proportional to the number of practicing teachers in that area. Potential backup teachers were also sampled, in case the initially selected teachers either refused or were unable to serve as judges. Sampled teachers were compensated for serving as judges, with the amount of compensation sufficient to reimburse their transportation costs and

the cost of hiring a substitute teacher for them. Although payments were made directly to the participating teachers, some of their school systems required the teachers to pay the cost of hiring the substitute teacher who assumed their teaching responsibilities on the day they participated in the standard-setting study.

How Many Judges Should be Used?

The simplest answer to this question is to use as many judges as possible, but that answer is not totally satisfying for two reasons. First, "as many as possible" might not be enough. Second, "as many as possible" might be far more than are needed, when compared to any reasonable criterion of sufficiency. So a more complex answer is needed.

If judgments could be obtained from an entire population of judges, the resulting test standard obviously would be more trustworthy than if judgments were obtained only from a small sample of judges. Some standard-setting procedures define as the final recommended test standard the mean of the standards recommended by sampled judges. For those procedures, the mean that would result from averaging the recommendations of all judges in an appropriate population can be regarded as a parameter the standard-setters would like to know. The desired population mean is estimated by the mean that results from averaging the recommendations of the sample of judges. The sample should be large enough that the estimated mean is not very different from the desired population mean. If the sample size is large enough that the sample mean would not fluctuate greatly among different samples of that size, it is reasonable to conclude that the mean recommended test standard produced by most samples will not differ much from the mean standard that would be recommended by the entire population of judges.

Fluctuation in the sample mean recommended standard is characterized by a statistic called *the standard error of the mean*. The standard error of the mean recommended test standard is computed by calculating the standard deviation of the test standards recommended by individual judges in the sample and then dividing that value by the square root of the number of judges in the sample. That is,

$$\text{Standard Error of the Mean} = \frac{\text{Standard Deviation}}{\sqrt{n}}$$

where n is the number of judges in the sample.

For standard-setting procedures that use the means of judges' individual recommendations as the final recommended standard, the difference between the recommendation produced by a sample of judges and the recommendation that would be produced by an entire population of judges (called the *estimation*

error) is unlikely to exceed two standard errors of the mean. More specifically, we can be about 95% confident that the estimation error will not be larger than two standard errors of the mean. The size of the sample of judges can be chosen so that this estimation error will be tolerably small.

For example, suppose that the standard deviation of the test standards recommended by 25 individual judges equaled 9 points. The standard error of the mean recommended test standard would be

$$\text{Standard Error of the Mean} = \frac{9}{\sqrt{25}} = \frac{9}{5} = 1.8 \text{ points.}$$

Two standard errors of the mean would thus be $(2)(1.8) = 3.6$ points, and we could be 95% confident that the estimation error would not be greater than 3.6 points. If we could tolerate an estimation error of 3.6 points, a sample size of 25 judges would be sufficient; otherwise, the sample size would have to be increased. If we wanted to be 95% confident that the estimation error was no more than 1 point, the standard error of the mean would have to be no more than 0.5 points. We could easily determine the size of the sample of judges needed to achieve this objective by solving the equation for the standard error of the mean for n , the one unknown value. From the definitional formula:

$$0.5 = \frac{9}{\sqrt{n}}$$

Rearranging terms and squaring both sides of the equation,

$$n = \left(\frac{9}{0.5} \right)^2 = 324.$$

so a sample of 324 judges would be required.

To compute the standard error of the mean recommended test standard, it is necessary to know the standard deviation of standards recommended by individual judges. Unfortunately, that value will not be known when a standard-setting study is being planned, and a required sample size cannot be computed with certainty. A standard deviation of 9 points was used in the preceding examples. This value approximates the average standard deviation of recommended test standards that resulted from use of Angoff's standard-setting procedure (described in the next section) with eight different tests in four empirical studies (Brennan & Lockwood, 1980; Cross, Impara, Frary, & Jaeger, 1984; Melican & Thomas, 1984; Mills & Melican, 1987). This very small data base provides support for the reasonable assumption that the standard deviation of judges' recommended test standards will be larger, the longer the test for which a

standard is sought; the correlation between test length and the standard deviation of recommended test standards was found to be 0.67.

It would be convenient if test length alone could be used to predict the standard deviation of recommended test standards that would result from use of the most popular standard-setting procedures. Unfortunately, few empirical studies include the standard deviation of recommended test standards, so a trustworthy prediction cannot be developed. However, the equation for Angoff's standard-setting procedure that resulted from the four studies involving eight tests was:

$$\text{Standard Deviation} = -7.36 + 0.15 (\text{Test Length})$$

This equation would predict a standard deviation of recommended test standards equal to 0.14 for a 50-item test, 3.89 for a 75-item test, 7.84 for a 100-item test, and 15.14 for a 150-item test. Mills and Melican (1987) reported standard deviations of recommended test standards equal to 0.79 and 2.08 for two 34-item tests, so the predicted values for shorter tests undoubtedly underestimate actual values by substantial amounts.

When deciding the size of the standard error of the mean recommended test standard that can be tolerated, it might be useful to consider the standard error of measurement of the test for which a standard is sought. If the standard error of the mean recommended test standard were no more than a fourth as large as the standard error of measurement of the test, it can be shown that the standard deviation of errors due to sampling of judges and the unreliability of the test would only be 3% larger than the standard error of measurement alone. Thus sampling error would not contribute substantially to overall error. If the standard error of the mean were equal to half the standard error of measurement of the test, the total standard error would be 12% larger. The equation used to compute these values is shown in a latter section of this chapter, titled "Adjustment for Both Types of Error."

Some standard-setting procedures use the median of the test standards recommended by individual judges as the final test standard. To determine the sample size required by those procedures, the standard error of the median recommended test standard would be substituted for the standard error of the mean in the formulas given earlier. The standard error of the median recommended test standard is calculated by first computing the standard error of the mean, and then multiplying that value by 1.253. That is,

$$\text{Standard Error of the Median} = 1.253 (\text{Standard Error of the Mean}).$$

What Should Judges be Asked to Do?

This question is another way of asking "What standard-setting procedures should be used?" Although dogmatic answers to the question can be formulated (some researchers have been quite assertive in promoting their favorite standard-setting procedure), the research base supporting the choice of a particular method is sparse and inconsistent.

The most comprehensive available review of standard-setting methods (Berk, 1986) lists 31 procedures. However, many of these procedures are minor variants on others, and the vast majority have not been applied in large-scale studies. The author will limit his description of standard-setting methods to five procedures, each of which has been used in at least two states to establish standards on teacher competency tests or statewide competency tests for students. The results of using these methods have also been compared in several empirical studies, so that their properties are reasonably well understood.

Angoff's Method. This standard-setting method, in its original form and in several variants, has been used more widely than almost all of its competitors. If standard-setting applications in the health professions are ignored, it is clearly the most widely used procedure. Angoff's method is one of several that ask judges to make judgments about test items and examinees' likely performances on those items, rather than asking them to make direct judgments about individual examinees. Four of the five methods that are discussed in this chapter fall into this category. The appeal of the general approach is the relative ease with which judgments can be elicited and the feasibility of using judges without personal knowledge of any of the prospective examinees.

Angoff's method requires that each judge consider each item on a test and estimate (1971, p. 515):

... the probability that the 'minimally acceptable' person would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score.

To apply the Angoff procedure, judges typically assemble at a single location for training, but then work independently. Judges are asked to conceptualize a minimally acceptable examinee. In the case of teacher-certification tests, judges would be asked to conceptualize a minimally skilled and knowledgeable teacher, in the domain assessed by the certification test. Livingston and Zieky (1982) suggest that judges be helped to define minimal acceptability by having them review the domain that the test is to assess and then participate in a discussion on what constitutes "borderline knowledge and skills." If the judges can agree on a level of knowledge and skill that distinguishes between examinees who are

acceptable and those who are not, Livingston and Zieky recommend that the distinction be recorded, together with examples of examinee ability that are judged to be just above, and just below, the threshold. As an example of this procedure, Livingston and Zieky considered a test that was designed to assess the reading-comprehension ability of high school students. To develop a common definition of minimal competence for this test, the authors suggested that judges be asked to reach agreement on whether a student who was minimally competent in reading comprehension must be able to "find specific information in a newspaper article, distinguish statements of fact from statements of opinion, recognize the main idea of a paragraph," and so on. If a procedure of the sort recommended by Livingston and Zieky is followed, it is essential that the knowledge and skills used to define a minimally acceptable examinee be selected from the domain assessed by the certification test.

After they have attempted to agree on the definition of a minimally acceptable examinee, each judge, working independently, considers the items on a test individually and predicts, for each item, the probability that a minimally acceptable examinee would be able to answer the test item correctly. As an example of the application of this procedure, consider the following instructions that were used to elicit judges' recommendations on a standard for the Reading Subtest of the National Teacher Examinations in a situation where the test was being used to screen applicants to teacher-education programs (Jaeger & Busch, 1984):

In this session we will ask you to make a judgment about each item in the Reading section of the Communication Skills Test of the National Teacher Examinations. The type of judgment is as follows:

You are to imagine a hypothetical teacher education applicant who has the minimum amount of knowledge and skill in the subject area covered by the Reading section of the Communication Skills Test, necessary to succeed in the approved State teacher education program with which you are most familiar. This hypothetical student has minimally acceptable, but not outstanding, skills and knowledge and might be what we call a low "C" student. You are to estimate, for each item in the test section, the probability or "chances" that this minimally competent applicant would be able to answer the test item correctly. Probabilities are expressed as fractions or decimals and represent the chances or likelihood that something is true. The probability can range from .00, "not possible," to 1.00, "certain." A probability of .50 has a 50% likelihood; that is, the probability of occurring is the same as the probability of not occurring.

Record your estimate of this probability on the JUDGMENT FORM FOR STANDARD SETTING for each item in the test section, by drawing a circle around the number which corresponds to the interval containing your estimate. For example, if you believe the probability for an item is 0.10 or less, circle the number "1" for that item. The "1" represents all probability estimates that are at least zero and no more than 0.10. The number "2" represents all probability estimates that are at least 0.11 and no more than 0.20, and so on. The number corresponds to the probability value at the upper end of the range it represents. So if you estimate the

probability value for an item to be 0.56, you would circle "6" for that item, since your estimate of 0.56 is between the probability values 0.51 and 0.60 represented by number "6."

Now please use the JUDGMENT FORM FOR STANDARD-SETTING to estimate an individual probability value for each test item. Do not be concerned about the number of times you circle the same numerical value. Just consider each test item independently. Be sure that you circle only one number for each test item. Remember that for each item, you are to imagine a hypothetical applicant who has the minimum amount of skill and knowledge in the subject-area covered by the Reading subtest, necessary to succeed in the approved State teacher education program with which you are most familiar. You are then to estimate the probability that this minimally competent applicant would be able to answer the test item correctly.

Note that these instructions to examinees presume that they will select their probability estimates from a set of 10 value ranges — zero to 0.10, 0.11 to 0.20, 0.21 to 0.30 . . . 0.91 to 1.00 — rather than estimating the exact probability for each item. This procedure permits the use of mark sense response forms that can be read by an optical scanning machine, thus greatly increasing the efficiency of processing the standard-setting data. If the sample of judges is large, the resulting increase in processing efficiency is important. Angoff originally proposed that each judge estimate a probability for each item, not that each judge select a probability for each item. Thus the procedure we have described is a variant of Angoff's procedure. It has been called the *modified Angoff procedure*, although it is but one of many modifications that have been used with some success.

Once judges' estimated probabilities for each test item have been determined, the sum of the probability estimates is calculated for each judge. The sum for an individual judge is regarded as that judge's recommended test standard. For example, if on a five-item test, a judge using the modified Angoff procedure were to select the values 6, 4, 4, 7 and 8, representing the probability ranges 0.51 to 0.60, 0.31 to 0.40, etc., the probabilities that judge estimated would be assumed to fall at the midpoints of the ranges he or she had selected: 0.555, 0.355, 0.355, 0.655 and 0.755, respectively. These values would be summed to yield the value 2.675 as the judge's recommended test standard.

The final step in Angoff's procedure is to average the recommended test standards across all judges in the sample. Once that average is computed, it is rounded to the nearest whole number and taken to be the test standard recommended by the entire sample of judges.

Nedelsky's Method. This standard-setting procedure was proposed by Nedelsky in 1954, and is the most venerable of the procedures that are widely used. It can only be used with tests composed of multiple-choice items.

In its operational procedures, Nedelsky's method is similar to Angoff's method. Judges are typically trained as part of a group, and then provide their judgments independently. Judges must conceptualize a minimally acceptable

examinee. The judgment task of the Nedelsky procedure focuses on test items rather than on individual examinees. The recommendations provided by individual judges are averaged, and the average is rounded to a whole number, to compute a recommended test standard.

The Nedelsky procedure differs from Angoff's procedure in the specific judgment task it imposes. Each judge is asked to consider each test item independently. When considering a test item, a judge must predict the number of response options a minimally acceptable examinee should be able to *eliminate* as being clearly incorrect.

Nedelsky defined a statistic that he called the *minimum pass level* (MPL) for each item. The MPL is based on the assumption that, when considering a test item, an examinee will eliminate the response options that he or she knows are incorrect and will then guess at random among the remaining response options. Consistent with that assumption, the MPL for an item equals the probability that a minimally competent examinee would get the item right by eliminating clearly wrong response options and guessing at random among the remaining options. For example, if a multiple-choice item has five response options and a judge decides that two of them are so obviously incorrect that even a minimally acceptable examinee should identify them as being wrong, three response options would remain. According to that judge, the MPL for the item would be $1/(5-2) = 1/3 = 0.33$, which is the probability that the minimally competent examinee would guess the correct answer from the three response options that remained, after he or she had eliminated the two that were obviously wrong.

For each judge, a recommended test standard is computed by summing the MPL values produced by that judge's recommendation, across all items on the test. The sum would be equal to the judge's estimate of the average score that minimally acceptable examinees would earn to the test, if the assumption of random selection among remaining item options were correct.

The initial recommended test standard produced by Nedelsky's method is simply the average of the recommended standards computed for all sampled judges, rounded to the nearest whole number.

Ebel's Method. This standard-setting procedure might be regarded as a variant of Angoff's method since the two methods are conceptually similar. However, the methods differ somewhat in their operational details and have been shown in a number of empirical studies to produce somewhat different recommended test standards.

Ebel's (1972) method requires the selection of a representative sample of judges and, like Angoff's method, is typically applied in a group setting, even though judges work independently after they have been instructed in the tasks they are to perform. A judge's first task in applying Ebel's method is to classify each item on the test into a category that defines the item's difficulty and relevance. Although other labels could be used, Ebel suggested the following three levels of item difficulty — *easy*, *medium*, and *hard* — and the following

four levels of item relevance — *essential, important, acceptable, and questionable*.

Once a judge has classified each test item into 1 of the 12 cells formed by this three-level by four-level categorization scheme, she or he must answer the following question for the collection of items in each cell (Livingston & Zieky, 1982, p. 25): "If a borderline test-taker had to answer a large number of questions like these, who percentage would he or she answer correctly?"

Once a judge answers that question for each of the 12 cells, the percentage he or she proposes for a cell is divided by 100 to convert it to a proportion, and the proportion is multiplied by the number of test items the judge allocated to that cell. The resulting products are summed across all 12 cells to compute that judge's recommended test standard. As in Angoff's and Nedelsky's methods, the final recommended test standard is computed by averaging the recommended test standards computed for individual judges, and rounding the average to the nearest whole number.

Jaeger's Method. This procedure requires the selection of samples of judges that represent at least one population. A recommended test standard is computed separately for each sample of judges. Judges are trained in a group setting and complete the standard-setting procedure in that setting, but record their judgments independently.

Each judge must answer the following yes-or-no question for each item in the test for which a standard is sought (Jaeger, 1989):

Should every examinee in the population of those who receive favorable action on the decision based on test performance (e.g., every student who is admitted to a teacher education program, or every candidate who is granted an initial teaching certificate) be able to answer the test item correctly?

An initial recommended standard for a judge is equal to the number of test items for which she or he answers "yes" to the question. On the other hand, the "yes" recommendations are counted, and the total count is regarded as the judge's recommended standard. An initial test standard for a sample of judges is established by computing the median of the recommended standards for individual judges in that sample.

Jaeger's procedure is iterative by design. Judges are given one or more opportunities to reconsider their initial answers to the yes-or-no question. Prior to reconsidering their initial answers, judges are given data on the some test performances of a sample of examinees of the type that would be affected by the test standard (e.g., candidates for initial-teacher certification or applicants to teacher-education programs), in addition to data on the recommended test standards that resulted from the answers provided by their fellow judges.

The types of test performance data used in Jaeger's procedure have varied in its several applications. In one statewide study, judges (1) made their yes-or-no item recommendations in an initial judgment session; (2) were provided a table

that showed the proportion of a sample of examinees that answered each test item correctly (called *item p-values*) and a graph that showed the proportion of their fellow judges who recommended a test standard equal to or less than each position score on the test; (3) reconsidered their initial recommendations by answering the yes-or-no question for each test item a second time; (4) were provided a graph that showed the percentage of examinees who would have failed the test during its immediate prior administration, had the test standard been set at each possible test score, and a graph that showed the proportion of their fellow judges who recommended a test standard equal to or less than each possible score on the test during the second judgment session; and (5) reconsidered their second-session recommendations by answering the yes-or-no question for each test item a third time.

In a more recent state-wide application of Jaeger's method, only two judgment sessions were used. Between the first and second sessions, judges were given a table of item *p-values*; a table that showed the percentage of examinees with total test scores in a restricted range of low values who answered each test item correctly (i.e., examinees who might be considered minimally acceptable); a graph that showed the percentage of examinees who would have failed the test during its immediate prior administration, had the test standard been set at each possible test score; and a graph that showed the proportion of their fellow judges who recommended a test standard equal to or less than each possible score on the test during the initial judgment session.

Jaeger (1982) recommends that more than one population of judges be sampled (e.g., teacher educators in colleges and universities and certified, practicing teachers) and that the final recommended test standard be the lowest among the median recommended standards produced by the single sample of judges selected from each of the various populations.

In addition to its iterative nature, a distinguishing characteristic of Jaeger's standard-setting method is that it does not require judges to conceptualize a minimally acceptable examinee. One could argue that this latter characteristic is the only distinguishing feature of Jaeger's method, as the procedures proposed by Angoff, Nedelsky, and Ebel could be modified to provide judges with information on examinees' typical test performances and with opportunities to reconsider their initial recommendations. Cross, et al. (1984) applied such modifications to Angoff's procedure and Nedelsky's procedure, and found that they generally increased the consistency of recommended test standards.

The Contrasting-Groups Method. In this method, proposed by Zieky and Livingston (1977), the principal focus of judgment is on the ability of individual examinees rather than on the characteristics of test items. An appropriate population of judges is defined and a representative sample of judges is selected from that population, just as in the Angoff, Nedelsky, and Ebel standard-setting methods described earlier. The first task of the sampled judges is to develop operational definitions of three categories of competence in the domain of knowledge and/or skill assessed by the test for which a standard is sought:

"adequate or competent," "borderline or marginal," and "inadequate or incompetent." In a particular application, such as recommending a standard for a test used to screen applicants to a teacher-education program, the three categories might be defined in terms of performance that is "definitely consistent with admission," "marginal or unconcerned about admission," or "definitely inconsistent with admission."

The next tasks of the sampled judges is to assign each member of a representative sample or subsample of the population of examinees to the three categories they have just defined. In other words, each judge must consider each sampled examinee and assign him or her to one of the categories "adequate or competent," "borderline or marginal," or "inadequate or incompetent." Clearly, judges can make these assignments only for examinees whose performances they know or about which they can learn. To avoid tautological assignment, in making their assignments, judges cannot use examinees' performances on the test for which a standard is sought.

Once all sampled examinees have been assigned to a category by each judge, the examinees take the test. The test standard that results from the contrasting-groups method is based on the test score distributions of examinees who have been assigned to the "adequate or competent" and the "inadequate or incompetent" categories. Hambleton and Eignor (1980) suggest that the two test score distributions be graphed together, and that the recommended test standard be the score value where the two distributions of scores intersect. They presume that the distribution of test scores of "competent" examinees will fall somewhat above the distribution of test scores of "incompetent" examinees, and that the two distributions will overlap. Livingston and Zieky (1982) suggest that the percent of examinees classified as "adequate or competent" be computed for each possible test score value, and that the recommended test standard be set at the score value where 50% of the examinees have been classified as "adequate or competent." If the examinee sample size is small, very few examinees will earn particular scores on the test, and the percentages required to recommend a standard with the contrasting-groups method will not be available or trustworthy. To address this problem, Livingston and Zieky recommend a procedure for smoothing or averaging the distribution of examinees' scores. The details can be found in their *Manual* (1982).

When used with tests for admission to a teacher-education program or for teacher certification, the contrasting-groups method imposes serious operational problems that might render it impractical. Most serious is the difficulty of finding judges who can knowledgeably classify individual examinees into one of the three categories. One might expect college and university professors to know some applicants for admission to a teacher-education program well enough to make such classifications, but their knowledge of the applicants might not be pertinent to the domain assessed by the admissions test. Also, professors would only be able to classify their own students and could not provide assessments for applicants from other institutions. To secure an adequate number of judgments, a

large sample of judges would be required, and the sample would have to be stratified across colleges and universities. Each judge would be asked to classify only the subsample of examinees he or she knew as a student. Although the contrasting-groups method is theoretically impervious to inconsistent classifications of the same examinees by different judges, it would break down in the face of many such inconsistencies. This problem could be eliminated, and the reliability of classifications could be increased, by having judges who knew the same applicants come to agreement on their classifications of such applicants.

The contrasting-group method appears to be totally impractical for setting a standard on a teacher certification test, because most, if not all, potential examinees would have completed a teacher-education program and would be judged competent to begin teaching by their college or university. A faculty member at such a college or university would be loathe to classify as incompetent an applicant who was approved for graduation from a teacher-education program by his or her institution. Even in the face of dissent among some judges, one could not expect many such classifications. The sample of "incompetent" examinees might therefore be so small that a reliable distribution of test scores for the incompetent group could not be obtained.

How the Methods Compare. Many researchers have compared the results obtained by using two or more standard-setting methods in similar circumstances. Much of this literature is summarized in Jaeger (1989). Although a complete, fully documented summary is beyond the scope of this chapter, a brief one is in order. Several researchers have found that Nedelsky's procedure produces a lower test standard than does Angoff's procedure, and that Jaeger's procedure produces a higher test standard than does Angoff's procedure. Few studies have compared Ebel's procedure or the contrasting-groups procedure with other methods. Angoff's procedure often, but not always, has been found to produce more consistent (and hence more reliable) test standards than many of its competitors. Because none of the procedures results in test standards that are consistently "best," and there are no criteria for identifying a "true" test standard, the use of several standard-setting procedures has been recommended by some practitioners. This recommendation is discussed in a later section of the chapter.

How Should Judges Be Trained?

The specific training that should be provided judges who participate in a standard-setting activity depends on the method used to recommend test standards. However, some general considerations apply to all methods. It should not be assumed that judges can do an adequate job of completing their standard-setting tasks in the absence of training. Judgment tasks should be described clearly and completely, and the potential use and interpretation of any data judges are provided should be explained and demonstrated. Simulation activities have been used quite effectively with some methods.

Most of the standard-setting methods described earlier require judges to consider the competence or ability of real or hypothetical examinees, either individually or collectively. However, only competence or ability in the domain assessed by the test for which a standard is sought should be considered. Judges must be trained to focus their judgment on the test domain and to avoid the temptation of making more global judgments. For example, if a test administered to prospective teacher-education students is in the domain of reading comprehension, judges must not define a minimally acceptable applicant as one who is adequately prepared in early childhood development.

The object of providing judges with data on the actual test performance of examinees or the recommendations made by their fellow judges is to help them become better informed and more reflective decision makers. To use data effectively, it is essential that judges understand them and also understand the formats in which the data are provided. It would be a mistake to assume that all judges, regardless of professional education or experience, will understand the meaning of test item *p*-values, or will know how to interpret a cumulative frequency distribution of test scores or recommended test standards. Specific training in the interpretation of such statistical displays, coupled with a guided simulation of the judgment activity in which such data are applied, has been found to be effective in at least one major standard-setting study (cf., Jaeger & Busch, 1984).

What Data Should Be Collected?

The data collected in a standard-setting study should be sufficient to permit decision makers to understand the "average" test standard recommended by professionals and the degree of consensus the professionals have achieved in providing a recommendation. If several types of judges have been used in the standard-setting study, the data collected should permit decision makers to consider the recommendations of each type of judge. For example, college faculty might recommend a different test standard than might practicing teachers. Or minority-group teachers might recommend a different test standard than might majority-group teachers.

To achieve these objectives, it is necessary to collect and record data on the test standards recommended by individual judges and to maintain records of such recommendations in a form that will permit summarization by type of judge as well as descriptive analyses of the recommendations provided by entire samples of judges.

How Should the Data Be Analyzed?

Those who proposed the standard-setting methods reviewed earlier included specific recommendations on the computation of test standards. The mean of the

standards recommended by individual judges is to be computed when the Angoff, Nedelsky, or Ebel methods are used. The median of the standards recommended by judges sampled from each population represented in the study is to be computed when the Jaeger method is used. The point that best distinguishes between the score distributions of "competent" and "incompetent" examinees is to be computed when the contrasting-groups method is used. However, each of these standard-setting methods admits computational variants.

It is well known that the mean of a score distribution is influenced far more by extreme scores, in this case the extreme recommendations of one or a few judges, than is the median. For this reason, it might be reasonable to use the median of the distribution of standards recommended by individual judges, regardless of the item-based standard-setting method employed. With any of the standard-setting methods described earlier, it is not unusual to find that one or more judges recommended a test standard that is equal to the maximum possible score on the test, or a score that is very close to zero. Although it would be tempting to eliminate such recommendations prior to computing a mean, censoring of the recommendations is done at some risk, and the point at which censoring should stop is unclear. It is also inconsistent and illogical to obtain carefully a representative sample of judges and then eliminate the recommendations of some of them, just because the recommendations are extreme. If a final recommended test standard is based on a median rather than a mean, the influence of extreme recommendations is mitigated without the need for censoring.

To inform decision makers about the degree of consensus (or lack of it) among judges, statistics other than measures of central tendency (mean or median) should be computed from distributions of recommended test standards. Some possibilities include the 10th, 25th, 75th, and 90th percentiles of the distribution of recommended test standards or a graphical display, such as a box-and-whisker chart, that portrays these statistics. The standard deviation of recommended test standards is not likely to convey meaning to decision makers, even though it is readily interpreted by statisticians and testing professionals.

As discussed in an earlier section of this chapter, the standard error of the mean (or median) recommended test standard provides important information about the stability of a recommended test standard, and is useful in determining whether a sample of judges is adequate in size.

Some Additional Matters of Judgment

The foregoing description of various standard-setting methods is limited to the features and procedures recommended by those who proposed the methods. Later researchers and users of these methods have proposed a variety of modifications and variants, intended to increase the reasonableness or the stability of the standards produced by the methods. Often, these variants are justified by an appeal to logic or judgment and are otherwise unsupported by theory or empiri-

cal findings. Several of the most important issues underlying proposed modifications are considered briefly in this section.

The Role of Test Results

Because the public conception of "standards" carries the connotation of absolute, rather than normative, performance expectations, some researchers have argued that data on the actual test performances of examinees have no legitimate role in a standard-setting procedure. Others have suggested that consideration of examinees' typical test performances will prevent judges from recommending standards that would produce unreasonable or infeasible policy outcomes — such as failing all applicants to teacher-education programs or granting initial teaching certificates only to a fourth of a state's teacher-education program graduates.

In keeping with the thesis that judges should anchor their recommendations in realistic information on the capabilities and performances of examinees who will be affected by a test standard, informing judges about the distribution of test scores likely to be achieved by typical examinees is recommended here. Judges will rightly differ in the extent to which their judgments are influenced by the distribution of examinees' typical test performances, but keeping them ignorant of such typical performances is inconsistent with the objective of having test standards recommended by knowledgeable, thoughtful professionals.

When judges are asked to estimate the item-by-item performances of minimally competent examinees, it is recommended that they be provided with data on the actual item performances of examinees, such as item p-values or with estimates of item p-values for a subsample of examinees with total test scores close to the test standard that is likely to be recommended. Cramer (1988) has proposed an iterative standard-setting procedure in which the judges' initially-recommended cut score is used to identify such a subsample of examinees. However, in at least one study (Jaeger, Cole, Irwin, & Pratto, 1980), judges reported that their recommendations were more strongly influenced by data on the proportion of examinees who would fail a test if the standard were set at each possible score value, than by data on the proportion of examinees who correctly answered each test item. It is not surprising that data on the proportion of examinees who would fail are influential, as they clearly convey the consequences of setting a test standard at various levels. This finding should not give comfort to those who use standard-setting procedures that require item-by-item judgments, however, because the processes judges use to consider such data when making recommendations concerning individual test items is not at all clear.

The Role of Several Judgment Opportunities

Among other researchers, Cross, Impara, Frary, and Jaeger (1984) found that the distribution of test standards recommended by individual judges became more homogeneous when the judges were given an opportunity to reconsider their initial recommendations. This finding applied to judges who used standard-setting methods proposed by Angoff, Nedelsky, and Jaeger. Having several judgment opportunities has produced inconsistent changes in the means of recommended test standards, with the mean increasing in some cases and decreasing in others.

Increasing the homogeneity of distributions of recommended test standards increases the reliability of the mean or median recommended test standard. Increasing homogeneity therefore provides sufficient empirical justification for giving judges an opportunity to reconsider their initial recommendations. In addition, having several opportunities for judgment is consistent with the goal of obtaining informed, well-considered recommendations.

Some would argue that the goal of increasing the homogeneity of recommended test standards is inappropriate or illusory for several reasons. First, variation in recommended standards can almost always be reduced by using fewer judges. Indeed, if a single judge is used, there will be no variation at all. Second, and of greater consequence, the fact that a sample of judges comes close to agreement on a single standard does not guarantee that such a standard is "correct" in any absolute sense. Thus, an increase in the reliability of a mean recommended standard should not be confused with an increase in the validity of the standard; a more reliable test standard is better only to the degree that it more closely approximates the recommendation that would be obtained from an infinitely large sample of judges.

Adjustment for Errors of Measurement

If the test standard is set at the value recommended by the selected sample of judges, half of the examinees whose true test scores are precisely equal to the test standard will fail the test, and half of the examinees whose true test scores are just below the test standard will pass the test. This is true because errors influence the observed value of an examinee's test score as well as the test standard recommended by a sample of judges. Because tests are less than perfectly reliable, an examinee's observed test score will fluctuate around the score that represents his or her true ability to perform in the subject area measured by the test. The difference between an examinee's observed test score and his or her true ability is called *measurement error*, and it occurs because of such influences as the sample of items that appear on the test, the way the examinee happened to feel at the time the test was administered, the peculiarities of the conditions under which the test was administered, and the success with

which the examinee guessed the answers to test items that he or she did not know. All of these factors influence the examinee's observed test score, but not his or her true performance.

Classical measurement theory assumes that, at any given test administration, half the examinees will have observed scores that contain positive measurement errors (and will thus exceed their true scores), and half will have observed scores that contain negative measurement errors (and will thus be exceeded by their true scores). It is also assumed that measurement errors are normally distributed. The standard deviation of measurement errors is a widely used index of the expected variation in test scores that is solely attributable to measurement errors. It is called the *standard error of measurement*.

To ensure that examinees with true levels of performance equal to, or slightly above, the recommended test standard do not fail the test just because of errors of measurement, the recommended standard can be adjusted by subtracting one or more standard errors of measurement. If the test standard were reduced by one standard error of measurement, 16% of the examinees with true performance *exactly equal to* the test standard would be expected to fail because of measurement error. If the test standard were reduced by two standard errors of measurement, 2% of such examinees would be expected to fail just because of measurement error.

Conversely, if the recommended test standard were increased by one standard error of measurement, 16% of examinees with true performance just below the test standard would be expected to pass the test, simply because of errors of measurement. And if the test standard were increased by two standard errors of measurement, 2% of such examinees would be expected to pass just because of measurement error.

Whether the recommended test standard should be reduced to protect truly qualified examinees from failure attributable to errors of measurement (called *false-negative* errors) or increased to protect society from truly unqualified examinees who would pass because of errors of measurement (called *false-positive* errors) is a matter of judgment that is best left to those responsible for setting test standards. However, it is the responsibility of those who recommended a test standard to include a recommendation on adjustment of that standard, to account for errors of measurement. The trade-off between false-negative and false-positive errors is an important matter of judgment. In the case of testing those who apply for admission to teacher-education programs, it might be argued that the societal risk associated with false-positive errors is low, as such errors are likely to be detected when unqualified students who are admitted because of measurement errors fail to perform satisfactorily. However, in the case of testing those who apply for teacher certification, the societal risks associated with false-positive errors (certifying persons who are unqualified to teach safely and effectively) is far greater, because such candidates' lack of qualification might not be detected immediately, if at all, and generations of students might suffer irreparable harm. Millman (in press) advances this argument and notes that the risk of

certifying unqualified candidates is exacerbated by current state policies that allow examinees who fail a teacher-certification test to attempt it again and again, until they ultimately pass the test or give up. Millman describes alternative passing rules that can be used to reduce the likelihood that unqualified examinees are granted admission, certification, or licensure, by capitalizing on measurement error and permissive policies that provide unlimited opportunities for retesting.

Adjustment for Sampling Error

The test standard recommended by a particular sample of judges will likely differ somewhat from the standard that would be recommended by another sample of judges or the standard that would be recommended if it were possible to secure judgments from all members of the population of judges. It might be considered unfair to fail examinees whose observed test score would equal the standard that would be recommended by the population of judges, just because their score is lower than the standard recommended by the selected sample of judges. Conversely, it might be considered unwise to pass examinees whose observed test score is equal to the standard recommended by the sampled judges, even though their score would be lower than the standard that would be recommended by the population of judges.

Just as it is possible to adjust a recommended test standard to account for errors of measurement, it is possible to adjust a standard to account for errors attributable to sampling of judges. If it is assumed that sampling errors average zero across samples of judges and that sampling errors are normally distributed, adjustments of the magnitudes discussed in the section on errors of measurement can be made to compensate for sampling errors. The only difference is that the recommended standard would be adjusted by adding or subtracting one or more standard errors of the mean recommended test standard, or one or more standard errors of the median recommended test standard.

If the procedure used to compute the recommended test standard is based on the mean of the standards recommended by individual judges in the sample, the adjustment would be made by adding or subtracting one or more standard errors of the mean recommended test standard. If the procedure used is based on the median of the standards recommended by individual judges, the adjustment would be made by adding or subtracting one or more standard errors of the median recommended test standard.

Adjustment for Both Types of Error

If it is desired to adjust the recommended test standard for errors of measurement and for sampling errors, this can be accomplished by adding or subtracting

one or more standard deviations that reflect both sources of error. Two steps are required to compute such a standard deviation. First, the Standard Error of Measurement is squared and then added to the square of the Standard Error of the Mean recommended test standard (or the Standard Error of the Median recommended test standard, as appropriate). Second, it is necessary to compute the square root of the sum just found. That is,

$$\text{Overall Standard Deviation} = \sqrt{(\text{Standard Error of Measurement})^2 + (\text{Standard Error of the Mean})^2}$$

When adjusting recommended test standards to compensate for potential errors of measurement or sampling error, the essential arbitrariness of such standards should be kept clearly in mind. Choice of a test standard is ultimately a matter of judgment, and its worth depends on the degree to which it produces reasonable results. Recommended test standards that are initially reasonable could be rendered unreasonable if adjustments for potential measurement error or sampling error are applied unthinkingly and without careful consideration of their consequences in terms of passing rates or failure rates. Although the use of sound, well-researched methods for eliciting judgments on appropriate test standards and adjusting initial standards is strongly recommended, the reasonableness and practicality of resulting test standards should also be investigated. The results of such investigations should be presented to policy makers whenever test standards are recommended.

Use of Multiple Standard-Setting Methods

Gallagher (1980) and Shepard (1980), among others, have suggested the use of more than one standard-setting method to satisfy a given standard-setting requirement. Their rationale is that different methods typically produce different recommended test standards, that different methods are often based on differing rationales and perspectives, and that no one method has been shown to be consistently superior to others, either on theoretical or empirical bases. These observations clearly justify the recommendation that, whenever feasible, several standard-setting methods be used to establish a test standard.

As noted earlier, several studies have shown that some standard-setting procedures produce more-reliable standards than others (Brennan & Lockwood, 1980; Cross, et al., 1984). This suggests that standard-setting recommendations based on several methods be accompanied by estimates of the precision of each recommendation. A graph that displayed the confidence interval surrounding each recommended test standard would probably represent the reliability of each recommendation in a form that was easily understood by policy makers.

When several standard-setting methods are used and each produces a different recommended test standard, policy makers are faced with an obvious dilemma. Possible solutions include (1) averaging the recommendations produced

by the various methods, taking account of their differing reliabilities, or (2) selecting the procedure that appears to produce the most reasonable recommendation, based on the availability of applicants or candidates and the need for accepted or certified personnel. This latter suggestion might appear to be totally unspecific, but it underscores the ultimately subjective and political nature of standard setting. It is the duty of those who recommend test standards to be scientific about their work. But those who ultimately set test standards cannot ignore such issues as society's needs for high quality education and an adequate supply of teachers as they make their standard-setting decisions.

Documentation and Reporting

The 1985 *Standards for Educational and Psychological Testing* (APA) and the *Uniform Guidelines on Employee Selection Procedures* (EEOC) require that standard-setting methods be fully documented. Required documentation extends to a report on the specific procedures used, the standard error of measurement of the test for score levels at or near the test standard, the size and composition of samples of judges, and the qualifications of judges. Instructions and training given the judges should also be documented, as should the detailed results of the standard-setting activity and any data analysis procedures that involved censoring of data.

Where to Look for Additional Information

A number of research articles and studies on standard setting have been referenced earlier in this chapter. Useful reviews of the literature on standard setting and recommendations on operational practice can be found in a journal article by Lorrie Shepard (1980), a book chapter by Ronald Hambleton and Daniel Eignor (1980), a manual on standard setting by Samuel Livingston and Michael Zieky (1982), a review-journal article by Ronald Berk (1986), and a book chapter by Richard Jaeger (1989). Complete references to these sources and other work cited in the chapter follow.

Note

1. The word *arbitrary* is used here in accordance with the second meaning provided in the *Oxford English Dictionary* (1971): "Relating to, or dependent on, the discretion of an arbiter, arbitrator, or other legally-recognized authority; discretionary, not fixed (p. 107)." Even though the judges used to recommend test standards rarely have legal authority to do so, those who set standards on teacher-certification tests typically have statutory authority to establish education policy. The word *capricious* is also used in accordance with the second meaning provided in the *Oxford English Dictionary* (1971): "Full of, subject to, or characterized by caprice; guided by whim or fancy rather than by judgment or

settled purpose, whimsical, humourous (p. 335)." Although the words *arbitrary* and *capricious* are sometimes used interchangeably, their connotative meanings are quite different. Reflective consideration often underlies arbitrariness, but never capriciousness. This distinction is central to the contrast made here.

References

- American Psychological Association, American Educational Research Association, National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: Author.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137-172.
- Brennan, R. D. & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, 4, 219-240.
- Cramer, S. (1988, July). Unpublished internal memorandum. Atlanta, GA: Georgia Department of Education.
- Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement*, 21, 113-130.
- Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register*, 43, (166), 38290-38315.
- Folks, J. (1985). Stakeholders in teacher certification testing. In W. P. Gorth & M. L. Chernoff (Eds.), *Testing for teacher certification* (pp. 47-57). Hillsdale, NJ: Lawrence Erlbaum.
- Gallagher, J. J. (1980). Setting standards for minimum competency: A case study. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences*. Berkeley, CA: McCutchan.
- Gifford, B. R. (1987). Excellence and equity in teacher competency testing: A policy perspective. In M. L. Chernoff, P. M. Nassif, & W. P. Gorth (Eds.), *The validity issue: What should teacher certification tests measure?* (pp. 51-74). Hillsdale, NJ: Lawrence Erlbaum.
- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.
- Hambleton, R. K. & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences*. Berkeley, CA: McCutchan.
- Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis*, 4, 461-475.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (4th ed.). New York: American Council on Education Macmillan.
- Jaeger, R. M. & Busch, J. C. (1984, March). A validation and standard-setting study of the general knowledge and communication skills tests of the national teacher examinations. Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina at Greensboro.
- Jaeger, R. M., Cole, J., Irwin, D. M., & Pratto, D. J. (1980, May). *Recommendations of registered voters, high school teachers, high school counselors, and high school principals on passing scores for the North Carolina high school competency tests* (Form B). Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina at Greensboro.
- Linn, R. L., Madaus, G. & Pedulla, J. (1982). Minimum competency testing: Cautions on the state of the art. *American Journal of Education*, (November), 1-35.
- Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Melican, G. C. & Thomas, N. (1984, April). *Identification of items that are hard to rate accurately using Angoff's standard setting method*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Millman, J. (in press). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher*.
- Mills, C. N. & Melican, G. C. (1987). *A preliminary investigation of three compromise methods for establishing cut-off scores*. Research Report RR-87-14. Princeton, NJ: Educational Testing Service.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.
- Oxford English Dictionary (1971). *The compact edition of the Oxford English Dictionary*. Oxford, England: Oxford University Press.
- Rudner, L. M. (1987). *What's happening in teacher testing: An analysis of state teacher testing practices*. Washington, DC: U.S. Government Printing Office.
- Shepard, L. A. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement*. (pp. 59-71). Washington, DC: National Council on Measurement in Education.
- Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, 4, 447-467.
- Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. Baltimore, MD: Johns Hopkins University Press.
- Shepard, L. A., Kreitzer, A. E. & Graue, M. E. (1987). *A case study of the Texas teacher test: Technical report*. Los Angeles, CA: Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.
- Zieky, M. J. & Livingston, S. A. (1977). *Manual for setting standards on the basic skills assessment tests*. Princeton, NJ: Educational Testing Service.

19. Combining Evaluation Data from Multiple Sources

William A. Mehrens

Evaluations of elementary and secondary school teachers or prospective teachers are likely to lead to high-stakes decisions — at least in the minds of those being evaluated. While all decisions should be based on the best data that can reasonably be provided, high-stakes decisions place greater demands on the quality of the data. In general, the more data that are gathered, the better the decision is likely to be. Certainly it is conventional psychometric wisdom that one should use more than one piece of data as the basis for important decisions.

However, Sechrest (1963) has argued that additional data should provide new and useful information. The result of using the additional data should be an increase in the confidence of the inferences made from the data (incremental validity). He suggested that additional data may not even result in better decisions. Further, one must consider the cost of gathering data relative to the improvement in the decisions that may result from the data (Cronbach and Gleser, 1965). Gottfredson and Crouse (1986) refer to this as a trade-off between utility and validity. One of the major feasibility standards listed in the *Standards for Evaluation of Educational Personnel* is fiscal viability (Stufflebeam & Brethower, 1987). Gathering more data will not always result in better decisions because the new data may be redundant information or simply add more noise to the system than they contribute valid information. Certainly more data will not always be cost effective.

Although the topics of incremental validity, the validity-utility trade-off, and cost effectiveness are all important topics to be considered in determining what data to gather, the focus of this paper is on how, or whether, to combine multiple sources of data to make decisions. In general, the decision(s) to be reached from the data determine whether a composite (combined) score is necessary. Further, even if one desires a composite score, it may still be wise to report out the separate subscores.

AUTHOR'S NOTE: Acknowledgements are gratefully extended to Robert Linn and W. James Popham, who provided detailed and helpful comments on both the outline and first draft of this chapter. I also thank my colleagues Robert Craig, Robert Floden, Irvin Lehmann, Susan Phillips, and Herbert Rudman at Michigan State, who provided me with helpful comments on the first draft.

This chapter will cover briefly how the purposes of evaluation relate to the issue of whether (or how) to combine data; describe some data-combination models, their assumptions, and their advantages and disadvantages; discuss the likely impact of the combination model used on the decision reached; and briefly discuss two empirical and Monte-Carlo investigations of data-combination methods.

Purposes of Evaluation

Chapters 5 through 12 of this Handbook discuss several different purposes of evaluation. Because the different purposes require different decisions, the need for combining the data differs. For example, in assisting beginning teachers (chapter 8) and in evaluation for improvement/professional development (chapter 9) one wishes to obtain a profile of a teacher's skills for the purpose of professional development. In such cases there is no necessity to obtain an overall composite score of a teacher's skills. However, if several pieces of data all measure the same specific skill, those measures may need to be combined in some fashion in order to obtain a sufficiently reliable measure for making a decision regarding the need for development in that particular area. In evaluation for selection, licensure, tenure, or merit pay (chapters 6, 7, 10, and 11) a single decision is reached. For those purposes, a decision maker must either explicitly or implicitly combine the data in some fashion for each person being evaluated. However, for some decisions (e.g., licensure) one is interested in determining only whether an individual is sufficiently competent while for other decisions (e.g., merit pay or career ladder) one may wish to place the individual teacher's competence along a continuum. This distinction may impact on the relative efficacy of different combination models.

Schmidt and Kaplan (1971) thoroughly analyzed the long-standing controversy in industrial psychology of whether one should use a weighted composition criterion or multiple criteria. The composite approach argument is an economic value-to-the-organization one, and single decisions about individuals (e.g., raises) must be based on some combination of the data. Those advocating multiple criteria use a behavioral (psychological) argument and submit that if criterion elements have low (or zero) correlations with one another:

... then they are obviously measuring different variables and weighting them into a composite results in scores that are so ambiguous as to be uninterpretable (Schmidt & Kaplan, 1971, p. 420).

Schmidt and Kaplan conclude their paper by suggesting that practicing industrial psychologists are interested in both economic and psychological ends. They should obtain a composite representing an economic construct to achieve practical goals but also analyze the separate elements to achieve psychological under-

standing. I submit that both of those goals are also appropriate for various teacher-assessment decisions.

Keeping in mind the relationship between purposes and methods, let us consider some data combination methods.

Data Combination Methods

Clinical Versus Statistical

One common distinction among data combination methods is whether the data are combined in a clinical or statistical fashion. Combining data in a clinical fashion is akin to eyeballing the various pieces of data to arrive at a final score on decision. Statistical weighting is the use of fixed weights for the data in a mathematical equation. Note that in comparing data combination approaches, the distinction is *not* on the type or amount of data. One has not compared the data combination methods unless the *same data* are used. Research generally has shown that when a criterion measure is available (allowing an optimal statistical combination to be determined via an empirical procedure), then statistical prediction provides greater accuracy than clinical prediction (see Dawes & Corrigan, 1974; Meehl, 1954). Data are more sparse in comparing logically derived algorithms with the clinical approach although even arbitrary statistical weighting generally out performs clinical (Nisbett & Ross, 1980).

Thus, if one has numerical values for a variety of measures of teacher quality, it is preferable to combine them in some algorithmic fashion rather than allowing some one to eyeball them to form some conclusion about the quality of the teacher. This is not to suggest that it is inappropriate to arrive at the numbers through a judgmental (clinical) approach. Indeed, there are advantages to having professionals make professional judgments about the quality of a teacher with respect to some skills rather than to simply tally up the "champagne bubbles" (Popham, 1987). However, once the judgment about a particular skill has been made, it should be quantified if the judgment about that skill is to be combined with other data. If one desires a fair evaluation, it should not be left to the idiosyncratic judgments of various evaluators to determine how much to weight each variable.

Conjunctive, Disjunctive, and Compensatory Models

Three common statistical decision-making models for combining data are the conjunctive model, the disjunctive model, and the compensatory model (see Gulliksen, 1950). The conjunctive model is frequently referred to as a multiple-cutoff model. In the conjunctive model, individuals must pass all measures to be acceptable. In the disjunctive model, individuals need only to be found acceptable on one of the measures. In the compensatory model, one combines the data

in some, typically additive, algorithm that allows low scores on score variables to be compensated for by high scores on other variables. For some situations, it is appropriate to combine the conjunctive and compensatory approaches. That is, it is appropriate within limits to allow one variable to compensate for another one.

The disjunctive model typically is used only when there are multiple measures of the same variable and one wants the individuals being evaluated to benefit from remediation or positive errors of measurement, but not to suffer from negative errors of measurement. Thus, for example, in a licensure test, applicants have several chances to retake the test. If on any occasion the applicant passes (even though it may be because of positive errors of measurement), he or she is considered acceptable.

Psychometric Issues. The major issue to be considered in choosing between the conjunctive and compensatory models is whether a surplus of one variable can compensate for a deficiency in another. Because compensation cannot possibly occur unless there is a positive monotonic relationship between the dependent variables and the independent variable (i.e., no point on the regression line with negative slope), the related psychometric issue is whether the relationships between the dependent measures and the criterion measure are linear (or at least positive monotonic). Also related are whether one is attempting to predict some criterion measure, and whether that measure exists in the real world or only theoretically.

Compensation or Noncompensation. It is not always true that an excess of one variable can compensate for another. Cronbach (1970) provides an example of recruits failing sonar school because of poor tonal judgment. They had been selected for the school because their high mechanical comprehension raised their mechanical tonal composite enough to conceal their tonal weaknesses.

In education, there are many possible areas where an excess of one characteristic will not compensate for a lack of another. For example, a high school mathematics teacher both needs to know some basic mathematics and needs to be able to relate to high school students. An excess of sensitivity to student needs cannot compensate for a deficiency in mathematics knowledge or vice versa. Thus, to be certified, a prospective teacher should be above some minimum score on both variables. The National Education Association (NEA) supports this notion:

A written test should be only one criterion for determining a student's readiness to practice the profession. The second part of the examination would measure the knowledge and skills of prospective teachers as demonstrated by their ability to perform effectively in the classroom. Newly credentialed teachers who have been successful on both parts of the exam should then begin an internship program. . . . (Cameron, 1985, p. 27).

However, for many educational decisions, it may make sense to allow a large amount of one variable to compensate for a small amount of another. This is particularly likely to be true within limits. Thus, one may wish to employ a combined conjunctive-compensatory approach. For example, a great deal of care in preparing high-quality lesson plans may compensate for some less than perfect skill at classroom control. If criterion data are not available, the beliefs of the evaluator will determine the choice of model.

Linearity. If one is making a predictive inference about a continuous variable, and if there exists a criterion measure so that one can validate the prediction, then there exists a common empirical approach to weighting the data in a compensatory fashion. This is the familiar multiple regression approach. (A discriminant function could be used if the criterion data were not continuous.) The model has been used extensively for selection decisions in employment and higher education. The equation assumes linearity of the predictors with the criterion measure although it may be useful if the relationships are only monotonically increasing (Dawes & Corrigan, 1974). The equation provides weights such that there is a maximum correlation between the weighted composite and the criterion measure for the specific sample of individuals whose scores were used in the regression equation. This approach has the advantage of being totally objective and providing an index for the effectiveness of the procedure (Horst, 1966). Because the equation maximizes the relationship for a particular sample, optimal weights for that sample will not be optimal for another sample. Research (Thorndike, 1986) has shown, for example, that with small samples, one gets a better cross-validated prediction of job success just using a measure of general ability (*g*) than using the subtest scores on a vocational aptitude battery. Although somewhat debatable, considerable evidence exists that, if one has converted all the data to a common metric (e.g., *z* or *T* scores), using unit weights — which are less susceptible to shrinkage in a new sample — is about as effective as employing the multiple regression equations (see Aamodt & Kimbrough, 1985; Schmidt, 1971; Wainer, 1976).

If a criterion variable does not exist in practice (though it may exist in theory) one cannot employ a purely empirical regression approach. However, one may still believe that an excess of one variable can compensate for a lack of another one. That may well be true if the variables are presumed to be theoretically related to a criterion measure in a monotonically increasing fashion. However, in such a case the weights will have to be determined in a logical rather than an empirical fashion. Such an approach is not totally objective with respect to assigning the weights, and one cannot obtain any measure of the effectiveness of the approach.

Gulliksen (1950, pp. 331-358) reviews a variety of weighting methods when no criterion is available. Darlington (1970) has proposed a system which depends upon *estimating* the relative sizes of the correlations between the predictors and the *unobserved* construct. Others (e.g., Fralick & Raju, 1982) have conducted studies comparing various weighting schemes. In general, both em-

pirical and theoretical literature suggests that there will be high correlations across the composites formed by different weighting schemes *if* the average intercorrelations of the scores to be combined and the average of the weights are both positive. However, if there are only a few scores with low intercorrelations and one compares alternate sets of weights with large variances and low intercorrelations, the weighting can make an appreciable difference (see Gulliksen, 1950, pp. 355-356 for appropriate formulas to obtain the correlation between two composites using different weights).

A major problem in the logical assignment of weights is that the variables are frequently not measured on the same metric (scale). It is difficult to know what the effect of weights are if the scales are not comparable. For example, to combine a principal's rating on a 5-point scale with a test score on a scale from 1 to 50 is as nonsensical as to average temperatures where some are measured on the Fahrenheit and some the Celsius scale. Thus, for logical weighting it is recommended that all data be converted to a common scale. Further, unless the variabilities of the measures are equal, the nominal (announced) weight will not be the same as the actual weight. Thus, it is a common practice to use a standardized scale such as a *z* score or a *T* score. This process equates the metrics both in terms of means and standard deviations and makes the nominal weights closer to the effective weights. (See Terwilliger and Anderson, 1969, for a discussion of cases where standardization may not have much of an impact.)

After the metric conversions, the variables should be weighted on their judged importance. One should keep in mind the known or expected reliability of the measures. In general, it makes sense to weight less reliable measures less heavily in the composite. Some psychometricians have suggested that the weight of a variable be its reliability, but that seems unwise because it ignores the judged relevance of the variable. It seems more reasonable to attenuate the judged relevance of the variable rather than the variable itself by the reliability.

If criterion data are unavailable, it seems useful to consider what is generally known about the variables related to job performance and employ validity generalization. Voluminous research by Hunter and Schmidt (see Hunter, 1986) and others suggests that general ability (*g*) is the major variable in predicting job performance. Thus, if one could factor the various independent measures and obtain their *g* loadings, one could use these loadings as weights in arriving at a compensatory composite. While this approach could clearly be defended based on the psychometric research on job performance, it presents some problems. First, it may not always be possible to extract a *g* factor from the scores. Even if a common factor emerges from a factor analysis of a number of measures of teacher competence, it may not reflect general ability.

If a general ability factor does emerge, three other problems exist with respect to the acceptability of weighting based on *g* loadings. First, there are many educators who insist that general ability is not that important for success in education. (These are probably people who do not read the industrial/organizational psychology literature and do not realize the deleterious inferences one

would be forced to draw about the teaching profession if they were correct about the lack of importance of cognitive ability.) Second, such weighting would likely increase the adverse impact of the decisions reached (Gottfredson, 1986). Third, such a weighting system would probably not be legally defensible for *licensure* decisions because it would be too difficult to defend both the content validity and the cutoff score.

Nonlinearity (Nonmonotone). If there is a nonlinear relationship between one predictor and the criterion measure (real or theoretical) it would not be optional to use a linear regression method above the point on the predictor variable where the slope of the curve changes. If the slope becomes zero, no compensatory model would be appropriate. As Mosier stated:

There are many psychological functions for which it is wholly unreasonable to assume a proportionate increase in the criterion with each unit increase in the measured function through the entire range of ability measured (1951, p. 790).

As Mehrens, Phillips, and Anderson stated:

If an excess of X_1 does not increase the amount of Y (or the assurance that Y will be above some cutoff Y_0), then X_1 surely could not compensate for deficiency in another variable (X_2) and therefore the conjunctive (multiple cutoff) model should be used (1987, p. 8).

It should be mentioned that Lord (1962, 1963) has demonstrated that even if the multiple cutoff approach is assumed to be appropriate, it is still not the best strategy to use with fallible (i.e., not totally reliable) data. However, the theoretically preferable contour lines approach produces a result that is not drastically different in effect from the less mathematically defensible — but far more understandable — multiple cutoff approach.

Nonpsychometric Issues. There are also some nonmeasurement considerations in choosing a model. First, it is typically easier for users of the data to understand how one arrived at a decision through a conjunctive model than a compensatory model. Second, more types of data can be gathered for the same cost (or, if no new types of data are gathered, there may be considerable savings in data-gathering expenses) if a sequential approach is used in the decision-making process.

Understanding. If the requirement for promotion, tenure, or whatever is based on a compensatory model score, it is difficult to communicate in advance of the evaluation just what the requirements are for each specific area. Theoretically, no level of performance on any of the areas is so low it would necessarily keep one from obtaining a sufficiently high composite score. If an individual receives a composite score based on a compensatory model it is difficult for the individual to use that information for improvement purposes. Even if each separate score is provided it may be very unclear to the individual which of the score(s)

contributing to the composite is (are) the one(s) in need of improvement. If the composite is five points below the cutoff needed for, say, the next step in the career ladder, which set of subareas are the ones which, if improved, are the most likely to result in an appropriately high composite? Of course, it is mathematically possible to compute how many z score points one would have to improve on any one subscore in order to achieve the necessary composite, but which of the subareas is most amenable to that improvement?

If the decision is based on a conjunctive model, it would be much easier to communicate in advance of the evaluation what the requirements are in each area. This information could be used by those individuals who are preparing to be evaluated. A conjunctive model is also easier to understand after the evaluation. If an individual does not qualify, specific areas of deficiency can be identified and remediated.

Further, it seems easier to communicate the meaning of the total set of requirements to the public if they are based on a conjunctive rather than a compensatory model.

Cost Savings — Sequential Decision Making. Note that there are three separate cut scores advocated in the quote given earlier by Cameron: (1) a knowledge test; (2) a performance measure; and (3) the successful completion of an internship. Note also in the quote that there is a sequential process to this conjunctive decision-making model. People are not allowed to begin an internship (the last cut score) until they have passed the previous two.

This notion of sequential decision making has been advocated for decades (Cronbach & Gleser, 1965; Guilford, 1956; Wald, 1950). It can result in considerable savings to the institution making the decision. This is true not only for hiring decisions — where, for example, many candidates are weeded out based on transcripts or letters of recommendation and thus only a portion of the applicants need to be interviewed — but also for decisions such as career-ladder placement. Consider a model where several variables (data-gathering procedures) are used for career-ladder placement. These might include years of experience, credits beyond the bachelors, quality of the teachers' lesson plans, student evaluations, and observations by outside evaluators. If one were employing a compensatory model, it would be necessary to gather all data for every applicant. However, in a conjunctive model, only those who passed the first four hurdles would need to be observed by the outside evaluators. This could conceivably result in considerable savings.

Unfortunately, some individuals misunderstand the sequential decision-making model. They think that when one uses such a model, the decision is made on only the final piece of data and this conflicts with the general notion that more than one piece of data should be used for important decisions. Of course, more than one piece of data is used in such a model. In fact, as Mehrens and Lehmann (1987) point out:

Actually sequential decision making usually results in *more* data being collected; one can often afford to gather *more* types of data for the same cost because not all data are gathered for all individuals (p. 95).

It is true that if one only considers those rejected, some get rejected on only one piece of data. In the example above, if three years of experience were set as a cut score, all those with less than three years experience would be rejected based on only that piece of data. Some of those individuals could conceivably be more deserving than individuals who eventually get the career ladder award. Nevertheless, it is common to make sequential decisions. For example, although some individuals without any formal training could conceivably be better kindergarten teachers than those with appropriate educational credentials, it would not be reasonable to interview *everyone* who happened to be interested in being hired — regardless of their formal training or their scores on other important variables. A sequential screening approach would be more appropriate although one may wish to use a compensatory model for the original sets of data to decide who to interview.

Which Method Should One Use?

Purposes

As discussed earlier, the purpose of the evaluation should influence the method of combining the data. If the evaluation is for the purpose of providing feedback to the individuals or the institution for purpose of improvement, it is probably not necessary to combine the data at all. Each important skill assessed should be reported out separately. For situations where it is necessary to make a single decision (e.g., selection, licensure, retention, merit pay), a single composite is obtained.

For selection decisions either in employment or higher education, it has been common practice to use the compensatory model. These are situations where there is a continuous criterion variable available and one wishes to maximize productivity. The assumption of linearity typically holds sufficiently well and one predictor variable can typically compensate for another.

For licensure decisions, the conjunctive model is the one most frequently used. The reasons are based on both the theoretical rationale behind the purpose of licensure and the practice of licensure. Theoretically, the purpose of licensure procedures is to exclude those who are not minimally qualified. The purpose is not to predict degree of success. Thus, the same inference (minimally qualified) is being drawn about a person who scores far above the minimum on a licensure process as for someone who scores just barely above the minimum cut score. This suggests that a licensure exam score is treated as nonmonotonic relative to the criterion.

Some have suggested that a compensatory model for licensure would be more appropriate and have suggested using college GPA along with the score on a licensure test. Mehrens, Phillips, and Anderson (1987) examined the logic and empirical basis for this suggestion and found it to be without merit. The lack of merit is attributable both to the theoretical rationale for licensure discussed in the previous paragraph and to the inadequacy of grades in measuring only relevant knowledge for licensure. Neither of these reasons points to a general deficiency in the compensatory model.

Comparing Methods

As Mehrens (1986) has pointed out, in any comparison of the models it is necessary to equate the number of individuals who succeed under each approach. Otherwise,

picking a model gets based on one's beliefs about the relative costs of false acceptances (positives) and false rejections (negatives) rather than on a preference for a data combination approach (p. 8).

Without a criterion measure, there is no totally psychometric way to determine which data combination model produces the "best" results. However, we do know that the methods will lead to more similar results in some situations than others. For single decisions (select/reject, licensure/don't, etc.) the differences in the conjunctive and compensatory models are greatest when the selection rate is close to 50% and the independent variables are not correlated. At the extremes — if all or none are selected, or if the variables are perfectly correlated — the two models lead to identical results.

Two Relevant Studies. Mehrens, Phillips, and Anderson (1987) investigated the use of GPA and tests in both a conjunctive and a compensatory model for teacher licensure decisions. As mentioned earlier, the logic of that situation supports a conjunctive model. Using a Monte-Carlo approach, the authors simulated the results for both models, allowing one retake on the licensure test. As they concluded:

The best decision rule for minimizing false negatives and total errors is to use graduation from an approved program to determine eligibility to take the licensure test, to allow for retakes, and to use zero weighting of the GPA with test score. That is, use a conjunctive model for combining GPA and test score. If one is primarily concerned with false positives one perhaps should not allow retakes (depending on the relative costs of the two errors) (p. 29).

Mehrens, McLarty, Rakow, and Phillips (in press) analyzed the data from eight different sources in the 1985-86 Tennessee Career Ladder Teacher Evaluation Systems. Tennessee used a combined compensatory/conjunctive model.

While the data were gathered sequentially (i.e., all data were not gathered at the same time), no interim decisions were reached. That is, regardless of scores on the data gathered first, all data were gathered on all individuals. The findings from the Mehrens et al. study suggested that less data could have been gathered without having any major impact on the decisions reached and that a purely conjunctive model with less total data could reproduce the actual decisions reached with considerable accuracy. One of the conclusions of the authors was that a conjunctive model for data combination should be considered for career-ladder evaluation systems.

Thus, two recent exploratory investigations lend support to considering the use of conjunctive models in licensure and career-ladder evaluations.

Summary and Conclusions

When more than one piece of data is gathered in a personnel evaluation, one must consider whether and/or how to combine the data. If the evaluation is for the purpose of professional development, it may *not* be necessary or even wise to combine the data. Each important skill should be measured and reported separately and the professional development should be individualized and skill specific. However, for evaluations that result in a single decision (selection, licensure, tenure or merit pay) the decision maker must either explicitly or implicitly combine the data in some fashion for each person being evaluated.

One can combine data either clinically or statistically. Generally statistical combinations give superior results. Statistical combination models include conjunctive, disjunctive, and compensatory models. Disjunctive models typically are used only when there are multiple measures of the same variable and one wants the individuals being evaluated to benefit from remediation or positive errors of measurement, but not to suffer from negative errors of measure. The major issues to consider in choosing between the conjunctive and compensatory models are the purpose of the evaluation, and whether an excess of one variable can compensate for a deficiency in another. If a criterion measure does not exist, the issue of whether variables compensate for each other must be decided at the logical level.

If the criterion measure does exist, and if the independent variables are linearly related to it, a multiple-regression (compensatory) approach provides optimal weights. Multiple regression often works reasonably well if the relationships are nonlinear but monotonic increasing. Even if a criterion measure does not exist the compensatory model is likely the preferred approach if logic suggests (as is often the case) that the variables can compensate for each other with respect to overall quality of performance.

If compensation (or monotonicity) does not exist, a conjunctive model may be the preferred approach. The conjunctive approach has two practical advantages: it is easier to understand and one can couple it with sequential decision making, which could result in gathering considerable more types of data (or

result in cost savings). A disadvantage of the conjunction approach is that one poor (unreliable or invalid) measure can invalidate the whole process more readily than would occur under the compensatory approach.

Two recent exploratory investigations suggest that the conjunctive model may be more appropriate for licensure decisions and that it is worth serious consideration in career-ladder decisions.

Finally, one should recognize that a combined conjunctive/compensatory approach may at times be the preferred approach. Perhaps *within limits* a larger amount of one variable can compensate for a smaller amount of a second one.

References

- Aamodt, M.G. & Kimbrough, W.U. (1985). Comparison of four methods of weighting multiple predictors. *Educational and Psychological Measurement*, 45, 477-582.
- Cameron, D. (1985). The NEA position on testing in-service teachers. *Educational Measurement: Issues and Practice*, 4, 3, 26-27.
- Cronbach, L.J. (1970). *Essentials of psychological testing*. (3rd. ed.). New York: Harper and Row.
- Cronbach, L.J. & Gleser, G.C. (1965). *Psychological tests and personnel decisions*. (2nd ed.). Urbana, IL: University of Illinois Press.
- Darlington, R.B. (1970). Some techniques for maximizing a test's validity when the criterion variable is unobserved. *Journal of Educational Measurement*, 7, 1-14.
- Dawes, R. M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95-106.
- Fralick, R.D. & Raju, N.S. (1982). Comparison of five methods for combining multiple criteria into a single composite. *Educational and Psychological Measurement*, 42, 823-827.
- Gottfredson, L.S. (1986). Societal consequences of the g factor in employment. *Journal of Vocational Behavior*, 29, 379-410.
- Gottfredson, L.S. & Crouse, J. (1986). Validity versus utility of mental tests: Example of the SAT. *Journal of Vocational Behavior*, 29, 363-378.
- Guilford, J.P. (1956). *Fundamental statistics in psychology and education*. (3rd ed.) New York: McGraw Hill.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.
- Horst, P. (1966). *Psychological measurement and prediction*. Belmont, CA: Wadsworth.
- Hunter, J.E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29, 340-362.
- Lord, F.M. (1963). Cutting scores and errors of measurement — a second case. *Educational and Psychological Measurement*, 23, 1, 63-68.
- Lord, F.M. (1962). Cutting scores and errors of measurement. *Psychometrika*, 27, 1, 19-30.
- Meehl, P.E. (1954). *Clinical vs. statistical prediction*. Minneapolis, MN: University of Minnesota Press.
- Mehrens, W.A. (1986). Measurement specialists: motive to achieve or motive to avoid failure? *Educational Measurement: Issues and Practice*, 5, 4, 5-10.
- Mehrens, W.A. & Lehmann, I.J. (1987). *Using standardized tests in education*. (4th ed.). New York: Longman.
- Mehrens, W.A., McLarty, J.R., Rakow, E.A., & Phillips, S.E. (in press). Fiscal viability, conjunctive and compensatory models, and career-ladder decisions: An empirical investigation. *Journal of Personnel Evaluation in Education*.
- Mehrens, W.A., Phillips, S.E., & Anderson, A.E. (April, 1987). *Conjunctive versus compensatory models for teacher licensure decisions: Monte Carlo and logical investigations*. A paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Mosier, C.E. (1951). Batteries and profiles. In E.F. Lindquist (Ed.), *Educational Measurement* (pp. 764-808). Washington, DC: American Council on Education.

- Nisbett, R.E. & Ross, L. (1980). *Human inference: Strategies and shortcomings of human judgment*. Englewood Cliffs, NJ: Prentice-Hall.
- Popham, W.J. (1987). The shortcomings of Champagne teacher evaluations. *Journal of Personnel Evaluation in Education*, 1, 25-28.
- Schmidt, F.L. (1971). The relative efficiency of regression and single unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 31, 699-714.
- Schmidt, F.L. & Kaplan, L.B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology*, 24, 419-434.
- Sechrest, L. (1963). Incremental validity: a recommendation. *Educational and Psychological Measurement*, 23, 153-158.
- Stufflebeam, D.L. & Brethower, D.M. (1987). Improving personnel evaluations through professional standards. *Journal of Personnel Evaluation in Education*, 1, 125-155.
- Terwilliger, J.S. & Anderson, D.H. (1969). An empirical study of the effects of standardizing scores in the formation of linear composites. *Journal of Educational Measurement*, 6, 145-154.
- Thorndike, R.L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior*, 29, 332-339.
- Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83, 213-217.
- Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.

PART III

Cross-Cutting Perspectives

20. Legal Issues Concerning Teacher Evaluation

Michael A. Rebell, Esq.

The relationship between the courts and the teacher-evaluation process seems puzzling. In the past, when teacher evaluations were conducted sporadically and were based on no discernible psychometric standards (Rebell, 1974; Popham, 1988, p. 269), the courts generally deferred to the expertise of school administrators and declined to second-guess their judgments. Currently, as the state of evaluation practice gives promise of significant improvement, the extent and the depth of judicial scrutiny appear to be on the rise.

Ironically, then, as the quality of evaluations improves, more of them seem likely to be struck down by the courts. There is, however, a method to (or at least an explanation for) this seeming judicial madness. In order to provide such an explain, it is necessary first to explore more deeply the apparent contradictions.

I. The Courts' Enhanced Role

A striking example of the courts' traditional deferential attitude toward teacher evaluation was the 1973 decision of the United States Court of Appeals for the Eighth Circuit in *Scheelhaase v. Woodbury* (1973). That case involved the dismissal of an experienced teacher whose contract had previously been renewed annually over a ten-year period. The stated reason for her termination was that "she was professionally incompetent as indicated by the low scholastic accomplishment of her students on the Iowa Test of Basic Skills and the Iowa Test of Educational Development." Despite the fact that a number of expert witnesses testified that it was improper to use these test scores as a basis for evaluating a teacher's performance, the court dismissed Ms. Scheelhaase's claim. These expert views were considered largely irrelevant because "such matters as the competence of teachers and the standards of its measurement" are not matters of constitutional dimension.

AUTHOR'S NOTE: The author acknowledges with appreciation the insightful critiques of Donald Bersoff and Rogers Elliot, the research assistance of Anne Murdaugh, and the cite-checking of Adam Saravay.

The *Scheelhaase* case is noteworthy both because of the Court's apparent lack of concern with the serious psychometric issues raised by a reliance on student achievement scores as a sole stated basis for termination (Millman, 1981; Berk, 1988; Haertel, 1989; *compare* Calif. Educ. Code), and because of the extent to which the court accepted a school administrator's psychometrically unsubstantiated, and possibly erroneous, position. One of the concurring *Scheelhaase* case judges bluntly stated:

The Board was entitled to rely upon the recommendation of conclusions of its superintendent, notwithstanding the existence of strong opinions contrary to his regarding the use of the ITBS or ITED tests as a tool of teacher evaluation. . . . Thus, its decision, even though premised upon an apparently erroneous 'expert opinion' cannot be faulted as arbitrary and capricious. The Board's mere mistake in judgment or in weighing the evidence does not demonstrate any violation of substantive due process. (Emphasis added).

Traditionally, most other courts tended to adopt a similar deferential stance in teacher evaluation cases, although most judges might not have articulated their premises quite so candidly.

In recent years, the federal courts (although, generally speaking, not the state courts) have taken a markedly different approach in testing cases. The change is well illustrated by the legal reaction to the St. Louis School District's 1985 decision to institute a system-wide faculty evaluation system which was based, to a significant degree, on student achievement gains. In light of the decision in *Scheelhaase* (which was rendered by the federal Appeals Court having jurisdiction over both Iowa and Missouri), one might have expected the teachers' attorneys to discourage a federal court challenge, especially as the use of student achievement data in St. Louis provided for consideration of "mitigating factors" and was psychometrically more sophisticated than the idiosyncratic use of student achievement data in the *Scheelhaase* case. Nevertheless, the St. Louis Teachers' union mounted a class action against the new evaluation procedures. Despite the *Scheelhaase* precedent, the District Court Judge in St. Louis denied defendants' motion to dismiss on this issue, thereby calling for a trial on plaintiffs' claim that use of student scores on the California Achievement Tests to rate teachers, a purpose from which they were neither designed nor validated, was arbitrary, capricious and unconstitutional. *St. Louis Teacher's Union v. Board of Education* (1987).¹

Why had the judicial climate changed so markedly in the 15 years since *Scheelhaase* was decided? Four major developments appear to be relevant. First, wider use of improved evaluation techniques, in the wake of legislative reform initiatives, has led to a larger number of denials of certification and/or teacher terminations. Second, a disproportionate number of these teachers or would-be teachers are members of minority groups whose grievances are given special scrutiny by the courts. Third, legal developments have broadened the jurisdic-

tion, especially of the federal courts, to consider social reform issues, including testing matters. Finally, the judges' own experience in assessing psychometric techniques in employment discrimination cases in recent years, and their familiarity with educational reform issues from desegregation, special education and other school-based litigations, has made them both better able and more inclined to analyze educational testing challenges in-depth (Rebell, 1988).

In June 1988, the United States Supreme Court issued a ruling which is likely to accelerate the trend toward increased judicial involvement in teacher evaluation matters. That case, *Watson v. Fort Worth Bank and Trust* (1988), extended to judgmental employment practices the Court's 1971 holding in *Griggs v. Duke Power Company* (1971) that standardized employment tests having a disparate impact on minorities must be shown to be job-related. Although the Court's ruling in *Watson* was unanimous, there was substantial disagreement among the Justices as to how closely courts should scrutinize particular practices and validation techniques. Whatever the precise standard of review ultimately implemented, there is little doubt that the federal courts will be more likely to scrutinize nonobjective evaluation procedures as a result of *Watson*.

Increased judicial scrutiny of teacher evaluation practices raises questions that go beyond general concerns as to the legitimacy and capacity of court involvement in social reform issues (Rebell & Block, 1982). Because the state of the art concerning teacher-evaluation practices is at a sensitive developmental stage, extensive court intervention at this point can substantially influence — for better or worse — the future direction of basic practice in the field.

The noteworthy aspect of teacher assessments in this regard is that these evaluations serve a dual purpose. They are used both to promote teacher training and to rate individuals for termination (Wise, Darling-Hammon, McLaughlin, & Burnstein, 1984). There is an inherent conflict between these "formative" and "summative" functions, as the candor and cooperation necessary for staff development purposes is at odds with the self-protective, adversary mode of rating or dismissal decisions (Popham, 1988; Hunter, 1988). In the past, when summative decision making constituted only an extremely minor portion of evaluative activities,² the underlying conflicts rarely came to the fore. Since the adoption of reform legislation, evaluative techniques increasingly are being used to raise accountability standards by denying certification, retention or promotion to those who do not make the grade. Thus, summative decisions, the type of decisions which often lead to court cases, are becoming more significant.

The growth in the importance of summative evaluation decisions means not only that more cases will be brought before the courts, but also that awareness of adversariness may pervade the process and constrict the entire effort to improve the state of evaluative practice. It is widely acknowledged that there is no consensus on specific criteria for what constitutes good teaching (Wise, et al., 1984; Johnson, 1980). This means that effective evaluation necessarily must emphasize judgmental factors.

The institutional orientation of the courts, however, is to favor objectivity and procedural regularity, especially in areas where there is significant potential for subjectivity and bias. For this reason, it has been argued that legalization of the evaluation process may be inconsistent with the "natural systems model of school organization," which sees teaching as an art carried out in a "loosely coupled" organizational environment (Darling-Hammond, Wise, & Pease, 1983). In other words, although much of current theory of teacher evaluation emphasizes a need for sophisticated assessment techniques that keep faith with the integrity of teaching (Stufflebeam & Brethower, 1987; Bacharach, Conley, & Shedd, 1987; Tetenbaum & Mulken, 1988), actual practice may be pressed in a contrary direction if evaluation policy is heavily influenced by the courts, and courts do not take into account fully the realities of the school setting.

It is not clear, however, that increased judicial involvement will have such a detrimental impact. In some measurement situations, courts have exhibited a sophisticated understanding of the complex judgmental factors at stake, and their insistence on thorough-going implementation of improved, fairer assessment devices has enhanced, rather than impeded, the development of professional standards (Rebell, 1989). It is important that an effective dialogue take place among the courts, the legislatures, and the educational and measurement communities to maximize the possibilities of such a positive interrelationship in the rapidly emerging area of teacher evaluation.

For such a dialogue to occur, members of the educational and measurement communities must be aware of both the current state of the law and of emerging legal trends. The rest of this chapter will endeavor to meet that need, on an overview basis. The next section will discuss issues of initial certification, dismissal of probationary teachers, dismissal of tenured teachers and performance evaluations, which are largely based on state statutory standards and, therefore, mostly litigated in the state courts. Consideration will then be given to the constitutional implications of these issues, especially when they have a disproportionate adverse impact on members of minority groups. Such constitutional issues and discrimination claims tend to be brought in federal courts. Finally, the concluding section will return to a consideration of the relationship between the courts and improved teacher-evaluation practices.

II. State Law

Within the American federal system, education is primarily a state responsibility. The federal Constitution provides no explicit right to education (*San Antonio v. Rodriguez*, 1973), and the Tenth Amendment to the Constitution operates to reserve educational issues to the states. All 50 state constitutions provide, in one form or either, for the establishment of a state public education system. Typically, a state constitution will empower the legislature to establish school districts and to enact a statutory framework for the governance of educa-

tional affairs. The legislature may set forth detailed requirements concerning faculty hiring and evaluation policies in its statutes and codes, or it may delegate much of the authority for promulgating specific standards to the state commissioner of education or the state board of education.³

Most teacher evaluation issues that come before the courts involve questions of interpretation of the specific requirements of state statutes or state board regulations (which regulations generally have the "force of law"). Courts also rely on prior judicial decisions, both for precedents in interpreting particular statutory language and for "common-law" guidance on issues that are not specifically covered by statutes and regulations. Psychometric standards such as validity and reliability, as well as specific assessment techniques, are rarely incorporated in state laws, regulations, or common-law standards. Consequently, cases involving evaluation issues tend to focus on adherence to specific procedural requirements set forth in state law or on general common-law notions of fairness and "due process," rather than on professional psychometric standards.

A. Initial Certification

In every state of the union, persons wishing to teach in the public schools must be certified in accordance with specific statutory and regulatory standards (Stinnet, 1974). These laws tend to require a college degree, with minimum credit hours or courses in specified subjects, plus a showing of good personal character and, at times, passage of an examination.⁴ Exceptions to the specific course requirements are, however, granted in times of shortage, usually with designated on-the-job training requirements. See, e.g., 8 N.Y.C.R.R. /sec 80.18. Some states also allow exceptions for persons with special backgrounds or qualifications. See, e.g., 6 N.J. Admin. Code /sec 6:11-3.23. The authority to practice granted by a state certificate is usually limited to the area covered by the particular certificate, although these requirements tend not to be strictly enforced.

Until recently, many states effectively delegated responsibility for certification to teacher-training institutions by automatically granting a certificate to graduates of their state approved programs. In the wake of the 1980s reform movement, however, the overwhelming majority of states now require some type of written examination for entry into a teacher-education program or for certification after graduation (Darling-Hammond & Berry, 1988, p. 23).

A state certificate is a *sine qua non* entry requirement for employment in the public schools, but it does not entitle its holder to an actual job. Nor does it guarantee retention or renewal of the certificate. The state's authority over certificate issuance, renewal, and retention is continuing in nature. States may alter requirements from time to time, both with regard to the issuance of new certificates and the renewal or revocation of outstanding certificates (see, e.g., *Texas State Teacher's Ass'n v. State*, 1986 (teachers holding life certificates nevertheless required to pass examination as a condition of continued employ-

ment)). In rare cases, certificates may be revoked, but this requires a full hearing and a showing of statutorily defined cause (e.g., *Erb v. Iowa*, 1974 ([teacher's adultery held not to constitute sufficient basis for revocation of teaching certificate])).

Courts generally accord broad discretion to state certification officials concerning substantive certification standards. Thus, state officials are not required to demonstrate the job-relatedness of requirements for applicants for an elementary school teacher position to take foreign language courses or for would-be high school physics teachers to have studied the history of education.

Similarly, unless there is a showing of arbitrary action or abuse of discretion, the courts are not likely to overturn a licensing authority's decision to deny certification to an individual candidate who did not meet technical requirements, even if those requirements were never shown to have been validated (see, e.g., *Board v. Nyquist*, 1973 [court upholds denial of licensure to acting principal who had served in the position for ten years with positive evaluations but had failed written exam]; *Application of Bay*, 1963 [court upholds denial of certificate to applicant with 8-year old burglary conviction]). The type of situation a court will consider arbitrary is illustrated by a New York case where the court overturned the denial of a substitute teaching license on the basis of the applicant's excessive weight (*Parolisi v. Board*, 1967).

B. Dismissal of Probationary Teachers

Once a would-be teacher has obtained state certification, he or she may seek actual employment with a particular local school district. The initial appointment is normally for a probationary period of three years, although in some states the probationary period may be shorter (see, e.g., Md. Educ. Code Ann. /sec 6-202 [b]) (two-year probationary period), and in other states the period may be extended to allow a teacher an opportunity to improve unsatisfactory performance (see, e.g., Mo. Educ. Code /sec 168-221) (three-year probationary period may be extended six months). The purpose of the probationary period is to provide teachers an opportunity to learn their craft, and school districts an opportunity to assess teachers' competence, before a permanent hiring commitment is made. Accordingly, although there are exceptions based on particular statutory or contractual provisions, in most states, probationary teachers can be dismissed at the end of their initial term by a board of education for any reason and without any need to prove that they are incompetent or ineffective.³ Generally, the only procedural requirement is that notification of nonrenewal be provided on or before a specified date prior to the contract expiration time.

Courts tend to construe these notification requirements strictly, and, in many states, a failure to provide the requisite notice on a timely basis will result in the teacher acquiring automatic tenure by passage of time (see, e.g., *Mathews v. Nyquist*, 1979). Several states require a written statement of reasons and, upon the teacher's request, an opportunity for a conference before the school board.

Such differences are not, however, full evidentiary hearings which would require a school board to establish actual cause for the termination.

Consistent with the purpose of the probationary period and the lack of any legal requirement to show cause for nonrenewal, courts generally do not scrutinize closely the actual evaluation processes of a district that has decided to terminate a nontenured teacher's employment. Illustrative of the general practice is the case of *Kruse v. Board* (1975). There, a nontenured teacher claimed that the Board had failed to provide proper notice of its intent to terminate her contract and of her right to request a conference to discuss the reasons for her termination.

The court strictly construed the state's procedural requirements and, even though the Board had actually met with the teacher and had extensively discussed the issues, the Court ordered the Board to meet with her again, after giving the requisite full notice. The majority and concurring opinions emphasized, however, that the purpose of the statement of reasons and conference would be "to afford the teacher an opportunity to successfully argue [her] dismissal was arbitrary and capricious if she can prove that each of the stated reasons underlying [her] dismissal is trivial. . . ." In other words, although the Board was obligated to listen to the teacher's arguments, in the end it would make the final decision and would not need to justify the basis for that decision to the teacher or to a court.

C. Dismissal of Tenured Teachers

Teachers who are reappointed after the probationary period generally are entitled to permanent tenure in their position,⁶ subject only to removal for cause or for other statutorily permitted reasons such as necessary reductions in overall school staff.⁷ Thus, there is a presumption that teachers who have achieved tenure are competent employees and, therefore, school officials should not be permitted to dismiss them without establishing a strong justification for doing so. The causes for which a tenured teacher may be dismissed vary from state to state; upon a proper showing, they may include general categories such as "incompetence" (encompassing lack of knowledge, failure to adapt to new teaching methods, failure to maintain discipline, and excessive absence), "incapacity" (including physical or emotional illness), "insubordination," "unprofessional conduct" (including use of offensive language, use of corporal punishment, or involvement in shoplifting incidents) and "immorality" (including dishonesty, sexual misconduct, or criminal action) (Valente, 1987, pp. 210-216).

Tenured teachers are entitled to formal evidentiary hearings to establish the validity of the charges against them. The procedures for such hearings depend on particular state statutes or regulations, school board regulations, or union contracts. Generally speaking, the procedural due process concepts applicable to a tenure hearing would require reasonable notification of the charges, an opportunity for a fair hearing before an impartial tribunal, representation by counsel at

the hearing, an opportunity to examine adverse witnesses, and the right to appeal an adverse decision.

As was true with probationary teachers, statutory notice and other procedural requirements tend to be strictly construed (see, e.g., *Orth v. Phoenix*, 1980 [notification of inadequacy, sent by the principal rather than by the Board of Education, held not to satisfy the statutory notice requirement]). In contrast to the arbitrary and capricious standard applied to nontenured teachers, however, courts reviewing discharge hearings for tenured employees hold the school boards to higher legal standards, such as a showing of "competent and substantial evidence" or "a preponderance of the evidence" to support the dismissal (see, e.g., *Meredith v. Board*, 1974 [tenured teacher dismissal for teaching deficiencies and excessive absences upheld on the basis of "competent and substantial evidence" generated by 600 pages of hearing testimony]).

Although the "competent and substantial evidence" standard places a significant burden of proof on a school board, it still does not require a board to show that assessment devices used to justify a finding of incompetence were professionally validated. As the court explained in *Board of Fort Madison v. Youel* (1979), the preponderance of the evidence standard means that the board must list specific findings of fact on each of the major points to support its conclusions. These legal requirements may be met by setting forth fact-based conclusions reached in routine evidence procedures, and the courts are not likely to probe beyond the conclusions to consider the adequacy of the assessment criteria used or the qualifications of the assessors.

Dismissals for incompetency tend to be based on a number of factors or a pattern of behavior rather than on isolated incidents (see, e.g., *Mims v. West Baton Rouge*, 1975 [repeated violations of rules, lack of control of students and failure to follow administrators' instructions held to constitute incompetence]). Often courts will consider whether the teacher received adequate notice of a need to improve performance (see, e.g., *Aulwurm v. Board*, 1977 [teacher dismissal invalidated because of lack of warning of deficiencies required under state law]). Where conduct is patently unacceptable, however, a single instance without prior warning can constitute reasonable cause for dismissal (see, e.g., *Mott v. Endicott*, 1986 [teacher dismissed for striking student in the genitals]).

D. Performance Evaluation Laws

Approximately half the states have enacted laws governing teacher evaluation. These laws apply to a variety of employment decisions including retention, tenure, dismissal, salary, reassignment and reduction in force (McCarthy & Cambren-McCabe, 1987, p. 254). Their content and requirements vary substantially. For example, Iowa law notes only that the local school board must establish an evaluation system (Iowa Code Ann /sec 279.14). California, by way of contrast, requires each governing board to establish specific standards relating to: (1) pupil progress toward established achievement standards; (2) instructional

techniques and strategies; (3) adherence to curricular objectives; and (4) the establishment and maintenance of a suitable learning environment (Cal. Educ. Code /sec 44662). Florida specifically mandates the school board to provide professional training programs to "ensure that all individuals with evaluation responsibilities understand the proper use of the assessment criteria and procedure" (Fla. Educ. Code, /sec 231.29(2)). In Pennsylvania, employees must be evaluated "by an approved rating system which shall give due consideration to personality, preparation, technique and pupil reaction in accordance with standards and regulations for such scoring as determined by rating cards to be prepared by the Department of Public Education. . . ." (24 Penn. Stat. /sec 11-1123).

Although state courts will require strict adherence to the procedural aspects of these requirements (see, e.g., *Bd. of Pub. Educ. v. Pyle*, 1978) and will strike down an arbitrary failure to use any apparent evaluative criteria (see, e.g., *Hyde v. Wellpinet*, 1980 [court orders renewal of principal's contract in the absence of utilization of any specific evaluation standards]), the state courts tend not to probe the substance of the evaluative criteria or methods used. Here again, the courts tend to defer to the administrators' "expert" judgment and to accept the results of their evaluations without undertaking any independent analysis.

For example, in *Kudasik v. Board* (1983), a probationary teacher who had been appointed at midyear received an unsatisfactory rating at the end of the year. The teacher challenged the rating, and the termination decision based upon it, alleging lack of adequate orientation, discomfort with the school's team concept of teaching, and personal hostility toward her by the team leader. She further stated that she corrected many of the deficiencies that had been pointed out in the midterm observations and was told toward the end of the term that her progress was satisfactory. Although these allegations raised material issues of rater bias and conflict between formative and summative evaluative techniques, the court refused to enter into any substantive analysis of the actual evaluations. Since the school authorities had satisfied their burden of "going forward" with evidence of unsatisfactory ratings and adherence to statutory procedures, and the teacher had not shown any "fraudulent" or "arbitrary, capricious" actions, the rating and consequent dismissal were upheld.

Court decisions emphasize consistency and objectivity of procedures rather than scientific standards of validity. For example, in a tenure dismissal case, a Pennsylvania court cited as "a model" evaluation procedure the following:

The evaluations occurred at two levels. At the first level is the principal; if he rates a professional employee unsatisfactory, the matter is referred to the second level, the superintendent, for further evaluation. While a teacher might object to being rated so often in a short period of time by different persons, such a procedure is clearly in the employee's best interest since it brings into the evaluation different viewpoints, thereby lessening the influence personal bias and prejudice with respect to teaching methods can have. We are particularly impressed with the ratings

... [on] what was going on in the classroom at five minute intervals. This method of evaluation has given us the best picture of the learning atmosphere in a classroom that we have seen to date in an anecdotal record. (*Rosso v. Board*, 1977)

The fact that multiple levels of observation and assessment were undertaken in this case is impressive, especially as the observations were repeated over a two year period. Beyond the regularity of the procedure and the multifaceted sources of observational input, however, the court did not point to, let alone probe, the justification for any substantive criteria, nor did it require any specific level of training in evaluative techniques by the raters.

Thus, *Rosso's* "model" procedures and *Kudasik's* evidentiary standards illustrate both the strengths and the limitations of state court review of teacher performance assessment. Clear violations of procedural norms, strong indications of lack of consistency and objectivity in rating techniques, and blatant instances of arbitrary actions will be struck down. Generally speaking, however, state courts reviewing teacher evaluation practices will not analyze directly the substantive criteria used to evaluate teachers, nor the training or qualifications of the raters.

III. Federal Law

Although education is primarily a matter of state law jurisdiction, federal courts will intervene where actions of state or local officials contravene federal constitutional or statutory mandates. In regard to teacher evaluation, federal jurisdiction is invoked primarily in cases involving claims of discrimination under the equal protection clause of the Fourteenth Amendment, or under the antidiscrimination statutes enacted to protect members of racial and ethnic minorities, women, employees with handicapping conditions or older individuals. In addition, claims of unfair treatment because of exercise of First Amendment free speech rights, including union-organizing activities, or allegations of denial of Fourteenth Amendment rights to due process by tenured teachers or others with a reasonable expectation of continued employment will trigger federal court jurisdiction. When federal courts are presented with cases raising evaluation issues, their scrutiny tends to be more probing and thoroughgoing than that of the state courts, as will be illustrated in the subsections which follow.

A. Initial Certification

Currently, 41 states have mandated some form of standardized testing requirements as part of their certification process for prospective teachers (Darling-Hammond and Berry, 1988, p. 23). Because many of these exams have had a disproportionate adverse impact on minority candidates, teacher competency tests have triggered a number of large scale federal class suits.

The courts have not definitely resolved the technical question as to whether the employment discrimination strictures of Title VII of the 1964 Civil Rights Act apply to teacher certification cases (Rebell, 1986, pp. 380-381). Nevertheless, the courts clearly draw on their experience in applying the Guidelines for Employment Discrimination issued by the federal Equal Employment Opportunity Commission (EEOC) under Title VII, and the APA/NCME/AERA Standards for Educational and Psychological Testing which are specifically incorporated in the Guidelines, in their approaches to these cases (see, e.g., *United States v. Texas*, 1972; *United States v. South Carolina*, 1977). Thus, in contrast to the state courts' posture of deference to administrators' judgments, the federal courts, in cases where there has been a showing of disparate impact on minorities, shifts the burden of proof to the defendants and require them to show that their practices are "job-related" in accordance with professional validation standards (*Griggs v. Duke Power Company*, 1971; *Albermarle v. Moody*, 1974).

The first challenges to teacher certification tests occurred in the 1970s in the deep South where the National Teachers' Examination (NTE) was widely used as a certification or job retention requirement, with heavy adverse impact on black teachers in newly desegregated school systems. Since the NTE was designed to assess student knowledge of the typical curricula taught at teacher training institutions, rather than job-related skills correlated to the teaching situation in any particular state or local school district, the Educational Testing Service did not authorize the use of the NTE as a certification device. Under these conditions, the courts did not hesitate to enjoin the use of the NTE (see, e.g., *Walston v. Board*, 1973; *Georgia Association of Educators v. Nix*, 1976). Where the NTE had been validated, however (at least in relation to the curriculum being taught in teacher training institutions in the particular state), its use was upheld (*United States v. South Carolina*, 1977).

In contrast to the early NTE situations, current teacher competency tests generally are instituted only after extensive validation processes have been undertaken. Recent challenges to some of these tests have, therefore, involved thorough analyses of these validation techniques to determine whether they satisfy applicable professional standards. For example, in Alabama, teams of nationally known psychometric experts from throughout the country spent months studying all phases of the state's teacher testing program; at the trial, weeks of testimony and thousands of pages of exhibits challenging and defending all aspects of the test construction process were put into evidence.⁸

State education departments and testing companies have become increasingly aware of the fact that testing programs which deny certification to substantial numbers of candidates (and especially candidates from minority backgrounds) are likely to face court challenges. In the legal process, the testing programs will receive probing scrutiny, whether these challenges result in court decisions or in settlements. Testing companies and testing authorities are, therefore, giving more attention to the legal defensibility of their validation procedures. Such legal awareness is likely to result in greater compliance with established professional

and judicial standards, although strict compliance with legal mandates may not be synonymous with preferred professional practice (Madaus, 1987).

B. Teacher Dismissals

1. *Liberty and Property Interests.* In the absence of cognizable claims of discrimination or violation of First Amendment rights, federal courts will decline to review issues relating to a probationary teacher's nonrenewal (see, e.g., *Nicholson v. Board*, 1982). A teacher who has achieved tenure, or "de facto tenure" (see *Perry v. Sinderman*, 1972) under state law may, however, invoke federal justification if he has been denied basic procedural rights. Such a teacher is considered to have a protected "property" interest in continued employment or a "liberty" interest in his professional reputation,⁹ which entitles him to due process protections under the Fourteenth Amendment (*Board v. Roth*, 1972).

Due process is a constitutional concept of fundamental fairness. It does not define a fixed set of procedures applicable to all circumstances. Rather, due process entails a balancing of individual and governmental interests in each situation. As most state tenure statutes guarantee extensive procedural rights which meet federal constitutional standards, federal cases in this area tend to focus on whether the state procedures were fairly applied, or whether a teacher who does not technically have tenure nevertheless has a "property interest" or "liberty interest" that would entitle him or her to similar procedural rights.

Although one might expect that federal court review, to the extent that it enforces hearing requirements similar to state procedures, would be comparable to state court review, many aggrieved teachers assiduously seek federal court jurisdiction. This is probably because federal rules and court procedures, and the judges' lighter caseloads, tend to provide plaintiffs with more extensive opportunities to obtain access to school district records and to develop their evidence. Moreover, since federal judges hear a greater number of civil rights cases, they tend to examine administrators' motives and actions with a degree of practiced skepticism, especially in cases raising allegations of racial or sexual discrimination. Accordingly, review by a federal court will, in practice, tend to be more probing and demanding.¹⁰

2. *Free Speech Rights.* Traditionally, it had been assumed that school boards could dismiss teachers for expressing views considered objectionable by local administrators. In a major constitutional decision (*Pickering v. Board*, 1968), however, the United States Supreme Court held that public school teachers have a First Amendment right to speak freely on matters of public concern, even if their views are considered objectionable by their employers. In a later ruling which qualified the *Pickering* doctrine, the Supreme Court held that even though a teacher's right to free speech is constitutionally protected, a school board may, nevertheless, discharge an employee if the board can show that it had valid reasons for discharge independent of the expression at issue (*Mt. Healthy v. Doyle*, 1977).¹¹

The *Mt. Healthy* test has tended to expand federal jurisdiction over teacher dismissals since many discharged employees claim that they are suffering retaliation for exercising First Amendment rights whereupon school boards must show by a preponderance of the evidence that they had genuine independent reasons for terminating the teacher (see Frels, Cooper, & Reagan, 1984/1986, p. 3). In such a proceeding, the employee must first establish that the speech or activity at issue was constitutionally protected and was a "substantial or motivating" factor in the school board's adverse employment decision. If this is shown, the burden then shifts to the board to show independent grounds for the dismissal by a preponderance of the evidence.

The federal preponderance of the evidence standard tends to involve the court in a more substantial and more probing review of evaluation decisions than do the analogous state court evidentiary standards discussed in the previous section. This is because the First Amendment analysis necessarily requires a federal court to consider carefully the substance of charges of incompetence or insubordination in order to determine whether they are substantial or are mere "pretexts" asserted to justify an underlying retaliatory motive (see, e.g., *Cox v. Dardanelle*, 1986 [charges of failure to follow administrative directives held pretextual and insignificant]).

D. Performance Evaluation Standards

The federal courts have over the years considered performance assessment practices of private employers in the employment discrimination context. In *Albemarle v. Moody* (1974), the U.S. Supreme Court applied specific provisions of the EEOC Guidelines in invalidating subjective supervisor rankings in an industrial job setting. In cases decided since *Albemarle*, federal courts have invalidated not only subjective systems based on no written instructions (see, e.g., *Rowe v. GM*, 1972), but also more sophisticated "total assessment processes" which did not meet professional validation standards (see, e.g., *Fisher v. Proctor & Gamble*, 1970; see generally Ashe & McRae, 1985). The recent decision in *Watson v. Fort Worth Bank and Trust* (1988) will undoubtedly extend the federal courts' review to cover many more performance evaluation practices.

School-based performance evaluations have tended thus far to receive less extensive judicial scrutiny than performance evaluations in other contexts (see, e.g., *Davis v. Board*, 1979; *Zahorak v. Cornell*, 1984; *Holley & Field*, 1977). Whether this distinction will continue in light of *Watson* remains to be seen. The courts' past failure to scrutinize closely educational evaluations may have been part of a general judicial reluctance to deal with the complexities of performance assessment for "upper level jobs" (see, Bartholet, 1982). In some major class actions, however, courts have induced the parties to develop new performance assessment techniques as part of the remedy to replace a challenged evaluation system (see, e.g., *Chance v. Board*, 1972 [new on-the-job performance examina-

tion for school supervisors developed by experts hired by both parties and implemented after more than ten years of litigation]).

It should be noted that federal courts have required that "objective criteria" be used for any demotions, reassignments or dismissals in southern school systems undergoing desegregation. This stance resulted from the courts' skepticism concerning judgmental approaches implemented by school districts with long histories of intentional racial discrimination (*Singleton v. Jackson*, 1969). Under the *Singleton* standard, traditional rating sheets, which may well have been upheld under state court "substantial evidence" standards were struck down (e.g., *United States v. Texas*, 1972).

Conclusion

The foregoing overview of state and federal law applicable to teacher evaluations has shown a trend toward greater federal jurisdiction, as constitutional concepts and the coverage of federal antidiscrimination statutes tend to expand. Increased federal jurisdiction means not only that more courts will be looking over the shoulders of school evaluators, but also that judicial scrutiny is likely to be more probing.

The main area for extensive, broad-based federal court scrutiny thus far has been initial certification competency testing. It is not coincidental that this has been the area where the largest number of Black and Hispanic candidates were being denied access to the teaching profession because of testing requirements. Having found adverse impact on minorities, the courts scrutinized extensively the tests and the validation techniques at issue and applied professional psychometric standards in their analyses.

The current education reform movement has sought to promote higher levels of teacher performance by raising assessment standards not only at initial entry, but also on-the-job. Merit-pay increases, career-ladder promotions and minimum competency testing of incumbent employees are all aspects of the latter reform initiatives. In the near future, it must be assumed that teachers who rate poorly under these types of assessments will also seek judicial relief. Indeed, in the past few years, major class actions were initiated to challenge minimum competency testing programs for incumbent teachers in Arkansas, Georgia and Texas.

It is significant that all three of these recent challenges were settled or withdrawn after full implementation of the program showed that overwhelming majorities, in excess of 90% of all teachers, including minority teachers, passed the tests (although some individuals passed only after four or five attempts). There are two possible explanations for this phenomenon. One is that apprehension caused by the pendency of major law suits caused state officials to lower testing standards in order to insure high passing rates. The second explanation is that the testing programs (all of which included substantial remedial opportuni-

ties for teachers who initially failed) successfully integrated the formative and summative aspects of teacher evaluation.

To the extent that fear of judicial intervention caused a lowering of otherwise valid and appropriate cut scores, increased court involvement in evaluation matters is a worrisome prospect. If, however, awareness of legal accountability motivated state officials and test makers to build effective remedial mechanisms into the testing process, the incumbent competency testing cases would constitute instructive examples of the positive potential of judicial intervention on teacher evaluation practices.

In testing cases, the courts tend to press the parties involved to create acceptable remedies for practices which have been held illegal. Those remedies at times amount to grudging, minimal compliance with court mandates; at other times, they constitute creative new solutions to complex problems. Seeing the issue in this light, it becomes clear that the answer to whether increased judicial intervention will have a positive or negative impact on professional evaluation practice lies more with the educators and psychometricians than with the judges.

Notes

1. As of the date of this writing, no trial has been held and the parties have entered into serious settlement negotiations.

2. During a five-year period in the 1960s, for example, only 12 out of approximately 60,000 teachers in the New York City school system were dismissed based on charges of incompetence (Tractenberg, 1973, p. 29).

3. In states which authorize collective bargaining with teachers' unions, some aspects of the hiring and evaluation process may also be governed by union contracts. Usually these involve procedures for conducting evaluation and for filing grievances, but not substantive evaluative criteria.

4. In each state, the educational system, and the legal structure related to it, have emerged from a unique historical and political setting. For this reason generalizations about state law can only provide overview insights. Readers interested in particular issues should research the relevant laws in their jurisdictions.

5. Dismissal during a specified term of employment may, however, raise issues of breach of contract and may require a showing of adequate cause.

6. Some states entitle teachers to "continuing contract" rather than tenure status. Although not formally providing permanent employment, continuing contract statutes generally require a showing of cause and a full evidentiary hearing before a board can decline to renew a teacher's contract (see *Perry v. Sinderman*, 1972).

7. A number of state statutes allow for terminations related to declining enrollments, financial exigencies and school district consolidations. Usually, but not always, such reductions in force statutes require nontenured teachers to be laid off before tenured (Phay, 1980).

8. The court did not render a verdict in this case because a settlement between the parties was entered. The settlement contemplated development of a new test which would incorporate statistical techniques to minimize the use of particular items which have an adverse impact on black candidates (*Allen v. Alabama*, 1987; see generally *Rebell*, 1986, pp. 391-397). At the present time, however, it appears that Alabama has chosen to abandon its certification testing requirement altogether, rather than develop such a new test.

9. Theoretically, a probationary teacher might also have a liberty interest in her reputation, but courts rarely consider mere termination actions without public communication of damaging reasons, to create a stigma on a teacher's professional reputation, especially with nontenured personnel.

10. There are indications of new developments in constitutional law which may lead to even more intensive federal court scrutiny in teacher dismissal cases. At least one federal Circuit has begun to develop a "substantive due process right" of teachers to be free from arbitrary and capricious dismissals (*Moore v. Warwick*, 1986). The Court in *St. Louis Teachers Union v. Board* (1977) interpreted this right to require a full trial of plaintiffs' allegations of invalid and inappropriate testing practices (see also, *Plyler v. Doe*, 1982 ["middle ground" equal protection standard applied to education law case]).

11. Another 1983 Supreme Court decision, *Connick v. Myers* (1983), further qualified the *Pickering* doctrine by holding that an assertion of a right to free expression would not be upheld where the speech is part of a personal employment grievance.

References

BOOKS AND ARTICLES

- American Psychological Association, American Educational Research Association, National Council on Measurement in Education (1985). *Standards for educational and psychological testing*.
- Ashe, R. & McRae, G., (1985). Performance evaluations Go to court in 1980s, *Mercer Law Review*, 36, 887-905.
- Bacharach, S., Conley, S., Shedd, J., (1987). A career developmental framework for evaluating teachers as decision-makers, *Journal of Personnel Evaluation in Education*, 1, 181-194.
- Bartholet, E., (1982). Application of title VII to jobs in high places, *Harvard Law Review*, 95, 945-1027.
- Beckman, J., (1981). *Legal aspects of teacher evaluation*, Topeka, KS: National Organization on Legal Problems in Education.
- Berk, R. (1988). Limitations of using student achievement data for career ladder promotion and merit pay decisions. In Mitchell, J.V., Plake, B.S. & Wise, S.L., *Assessment of teaching*. Hillsdale, NJ: Erlbaum (forthcoming).
- Darling-Hammond, L. & Berry, B., (1988). The evolution of teacher policy. Santa Monica, CA: The RAND Corporation (March).
- Darling-Hammond, L., Wise, A., & Pease, S., (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, 53, 285-328.
- Delon, F. (1988). Assessment center screening for school administration certification and employment: Possible legal challenges, *Education Law Reporter*, 43, 841-848.
- Equal Employment Opportunity Commission (1978). Uniform guidelines on employee selection procedures, 29, C.F.R. pt. 1607. Washington, DC.
- Frels K., Cooper, T. & Reagan, B. (1984/1986). *Practical aspects of teacher evaluation*, Topeka, KS: National Organization on Legal Problems of Education.
- Haertel, E., (1989). Student achievement tests as tools of educational policy: Practices and consequences. In Gifford B., *Testing and the allocation of opportunity*, Vol. I, Boston, MA: Kluwer-Nighoff (forthcoming).
- Holley, W. & Field, H. (1977). The law and performance evaluation in education: A review of court cases and implications for use. *Journal of Law and Education*, 6, 427-448.
- Hunter, M. (1988). Reflecting a reconciliation between supervision and evaluation — A reply to Popham. *Journal of Personnel Evaluation in Education*, 1, 275-280.
- Johnson, S., (1980). Performance based staff lay-offs in the public schools: Implementation and outcomes, *Harvard Education Review*, 50, 214-233.
- Madaus, G. (1987). Legal and professional issues in teacher certification tests: A psychometric snark hunt. Lincoln, NE: Buros NEB Symposium paper.
- McCarthy, M. & Cambron-McCabe, N. (1987). *Public school law: Teachers and student's rights*, (2nd ed.). Newton, MA: Allyn & Bacon.

- Mehrens, W., (1987). Validity issues in teacher licensure tests, *Journal of Personnel Evaluation in Education*, 1, 195-229.
- Millman, J. (1981). Student achievement as a measure of teacher competence. In Millman, J., *Handbook of teacher evaluation*, (1st ed). Beverly Hills, CA: Sage Publications.
- National Commission on Excellence in Education (1983). *A Nation at risk: The imperative for educational reform*.
- Phay, R., (1980). Reduction in force and recommended policy. Topeka, KS: National Organizational on Legal Problems in Education.
- Popham, W.J. (1988). The dysfunctional marriage of formative and summative teacher evaluation, *Journal of Personnel Evaluation in Education*, 1, 269-274.
- Rebell, M. (1989). Testing, public policy and the courts. In Gifford B.R., *Testing and the allocation of opportunity*, Vol. I, Boston, MA: Keuwer-Nighoff (forthcoming).
- _____. (1988). Legal issues concerning bias in testing. In Allan R., Nassif P. & Elliot, M., *Bias issues in teacher certification issues*. Hillsdale, NJ: Erlbaum.
- _____. (1986). Disparate impact of teacher competency testing on minorities: Don't blame the test takers — or the tests, *Yale Law & Policy Review*, 4, 375-403.
- _____. (1974). *Teacher credentialing reform in New York state: A critique and a suggestion for new directions*, Lincoln, NE: Study Commission on Undergraduate Education and the Education of Teachers.
- Rebell, M. & Block, A. (1982). *Educational policy-making and the courts: An empirical study of judicial activism*. Chicago: University of Chicago Press.
- Rosenberger & Plimpton, (1975). Teacher incompetence and the courts. *Journal of Law and Education*, 4, 469-486.
- Stufflebeam, D. & Brethower, D., (1987). Improving personnel evaluations through professional standards. *Journal of Personnel Evaluation in Education*, 1, 125-155.
- Stinnet, T.M. (1974). *A manual on standards affecting school personnel in the United States*. Washington, DC: National Education Association.
- Strasle, (1985). Minimum competency testing and teachers for certification: Due process, equal protection and title VII implications. *Cornell Law Review*, 70, 494-524.
- Task Force on Teaching As A Profession (1986). *A nation prepared: Teachers for the twenty-first century*. New York: Carnegie Forum on Education and the Economy.
- Tettenbaum, T.J. & Mulken, T.A. (1988). Assessment of educational personnel and the twenty-first century. *Journal of Personnel Evaluation in Education*, 1, 235-240.
- Tractenberg, P., (1973). *Testing the teacher*, New York: Agathon Press.
- Valente, W. (1987). *Law in the schools*, (2nd ed.). Columbus, OH: Merrill Publishing Company.
- Weick, K. (1976). *Educational organizations as loosely coupled systems*. *Administrative Science Quarterly*, 21, 1-19.
- Wise, A., Darling-Hammond, L., McLaughlin, M., & Burnstein, H. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: The RAND Corporation.

STATUTES

- Cal. Educ. Code /sec 44662 [West supp. 1989]
- Fla. Educ. Code /sec 231.29(2) [1989]
- Iowa Code Ann. /sec 279.14 [West supp. 1988]
- Md. Educ. Code Ann. /sec 6-202(b) [1989]
- Mo. Educ. Code /sec 168.221 [Vernon supp. 1989]
- 6 N.J. Admin. Code /sec 6:11-3.23 [1986]
- 8 N.Y.C.R.R. /sec 80.18 [1987]
- 24 Penn. Stat. /sec 11-1123 [Purdon supp. 1989]

COURT CASES

A. State Cases

- Anderson v. San Mateo Community College Dist.*, 87 Cal. App.3d 441 (1978).
Application of Bay, 378 P.2d 558 (Ore. 1963).
Aulwurm v. Board of Education of Murphysboro Community Unit School Dist., 367 N.E.2d 1337 (Ill. 1977).
Board of Education v. Nyquist, 31 N.Y.12d 468, 341 N.Y.S.2d 441, 293 N.E.2d 819 (1973).
Board of Education of Fort Madison Community School District v. Youel, 282 N.W.2d 677 (Iowa 1979).
Board of Public Education v. Pyle, 390 A.2d 904 (Pa. Cmwlth., 1978).
Celestine v. Lafayette Parish School Board, 284 S.2d 650 (La. App. 1973).
Erb v. Iowa State Board of Public Instruction, 216 N.W.2d 339 (Iowa 1974).
Gabriel v. Trinity American School District, 350 A.2d 203 (Pa. Comm. Ct. 1987).
Holland v. Board of Education of Raleigh County, 327 S.E.2d 155 (W.Va. 1985).
Hyde v. Wellpinit School District No. 49, 611 P.2d 1388 (Wash. App. 1980).
Kruse v. Board of Directors of Lamoni Community School Dist., 231 N.W.2d 626 (Iowa 1975).
Kudasik v. Board of Directors, Port Allegany School Dist., 455 A.2d 261 (Pa. Cmwlth., 1983).
Mathews v. Nyquist, 412 N.Y.S.2d 501 (App. Div. 1979).
McCrum v. Board of Education of the New York City School District, 396 N.Y.S.2d 691 (App. Div. 1977).
Merideth v. Board of Education of Rockwood, 513 S.W.2d 740 (Mo. App. 1974).
Mims v. West Baton Rouge Parish School Board, 315 S.2d 349 (La. App. 1975).
Mott v. Endicott School District, 713 P.2d 98 (Wash. 1986).
Orik v. Phoenix Union High School System, 613 P.2d 311 (Ariz. App. 1980).
Parolisi v. Board of Examiners of City of New York, 285 N.Y.S.2d 936 (Sup. Ct. 1967).
Perez v. Commission on Professional Competence, 197 Cal. Rept. 390 (Cal. App. 1983).
Rosso v. Board of School Directors, 388 A.2d 1238 (Pa. Cmwlth., 1977).
Texas State Teachers Ass'n v. State, 711 S.W.2d 421 (Tex 1986).
Wheeler v. Yuma School District No. One, 750 P.2d 860 (Ariz., 1988).

B. Federal Cases

- Albermarle Paper Co. v. Moody*, 422 U.S. 405 (1974).
Allen v. Alabama State Board of Education, 816 F.2d 575 (11th Cir. 1987).
Board of Regents v. Roth, 408 U.S. 564 (1972).
Chance v. Board of Examiners, 458 F.2d 1167 (2nd Cir. 1972); 561 F.2d 1079 (1977).
Cleveland Board of Education v. Lauderhill, 470 U.S. 532 (1985).
Connick v. Myers, 461 U.S. 138 (1983).
Cox v. Dardanelle Public School District, 790 F.2d 668 (8th Cir. 1986).
Davis v. Board of School Commissioners of Mobile County, 600 F.2d 470 (5th Cir. 1979).
Faro v. New York University, 502 F.2d 1229 (2d Cir. 1974).
Fisher v. Proctor & Gamble Manufacturing Company, 613 F.2d 527 (5th Cir. 1970).
Georgia Association of Educators v. Nix, 407 F. Supp. 1102 (N.D. Ga. 1976).
Griggs v. Duke Power Company, 401 U.S. 424 (1971).
Harrison v. Lewis, 559 F. Supp. 943 (D.D.C. 1983).
Jarron v. Port Smith School District, 451 F.2d 1106 (1st Cir. 1971).
Johnson v. University of Pittsburgh, 359 F. Supp. 1002 (W.D. Pa. 1973).
Jones v. Jefferson Parish School Board, 533 F. Supp. 816 (E.D. La. 1982), *aff'd*, 688 F.2d 837 (5th Cir. 1983).
McEnteggar v. Cataldo, 451 F.2d 1108 (1st Cir. 1971), *cert. denied* 408 U.S. 943.
Moore v. Warwick Public School District No. 29, 794 F.2d 322 (8th Cir. 1986).
Mount Healthy City School District v. Doyle, 429 U.S. 274 (1977).
Nicholson v. Board of Education, 682 F.2d 858 (9th Cir. 1982).
Perry v. Sindermann, 408 U.S. 593 (1972).

- Pickens v. Oklahoma Municipal Separate School District*, 527 F.2d 358 (5th Cir. 1976).
Pickering v. Board of Education, 391 U.S. 563 (1968).
Pickering v. Doe, 457 U.S. 202 (1982), rehearing denied 458 U.S. 1131.
Powell v. Syracuse University, 580 F.2d 1150, 1153 (2d Cir. 1978), cert. denied 439 U.S. 984.
Rowe v. General Motors Corporation, 457 F.2d 348 (5th Cir. 1972).
St. Louis Teacher's Union, Local 420 v. Board of Educ. of the City of St. Louis, 652 F. Supp. 425 (E.D. Mo. 1987).
San Antonio School District v. Rodriguez, 411 U.S. 1 (1973), rehearing denied, 411 U.S. 959.
Scheelhaase v. Woodbury Central Community School District, 488 F.2d 237 (8th Cir. 1973), cert. denied, 417 U.S. 969.
Singleton v. Jackson Municipal Separate School System, 419 F.2d 1211 (5th Cir. 1969); reversed in part on other grounds, 396 U.S. 290 (1970).
United States v. South Carolina, 445 F. Supp. 1094 (D.S.C. 1977), aff'd 434 U.S. 1026 (1978).
United States v. Lulac, 793 F.2d 636 (5th Cir. 1986).
United States v. Texas Education Agency, 459 F.2d 600 (5th Cir. 1972).
Walston v. County School Board of Nansemond County, 492 F.2d 919 (4th Cir. 1973).
Watson v. Fort Worth Bank and Trust, 108 S. Ct. 2777 (1988).
Zahorak v. Cornell University, 729 F.2d 85 (2d Cir. 1984).

21. The Ethics of Educational Evaluation

Kenneth A. Strike

This paper has two sections. In the first, the author discusses some ethical concepts that concern teacher evaluation broadly. In the second, he deals with the question of legitimate authority over teacher evaluation.

I. Ethics and Evaluation

The author begins by sketching some general ethical principles. These general principles suggest abstract standards that the application of the more specific principles to be suggested shortly should satisfy.

The Principle of Benefit Maximization

The principle of benefit maximization¹ holds that the best decision is the one that results in the greatest benefit for the most people. Thus, it judges the morality of our actions by their consequences. The best action is the one with the best results. If, for example, the production of happiness is thought to be a basic good, then the principle of benefit maximization indicates that we should make those decisions and engage in those actions that result in the greatest happiness for the greatest number. This expression of the principle of benefit maximization is known as utilitarianism. (See Bentham, 1961; Mill, 1961.) The application of the principle of benefit maximization requires us to specify the goods we wish to maximize, identify a relevant population for maximization, specify what is to count as maximization and fully understand the consequences of our actions.²

The Principle of Equal Respect

The principle of equal respect requires that we respect the equal worth of moral agents. It involves three subsidiary ideas. First, the principle of equal

AUTHOR'S NOTE: I wish to thank Jonas Soltis and Adam Urbanski for their helpful comments on this paper.

respect requires us to treat people as ends rather than means. We may not treat others as though they were simply means to further our own goals. Second, a central fact about moral agents is that they are free and rational. Thus, we must respect their choices. Free people are entitled to their own concept of their own good. Third, as moral agents people are of equal worth. That people are of equal worth means that they are entitled to the same basic rights and that their interests are of equal value.

The principle of benefit maximization and equal respect for persons are not incompatible. However, they express two different views about what is central in ethics. The one taken to be primary will affect the interpretation and application of more specific ethical principles of the sort noted below. Generally, those for whom the principle of equal respect is central will treat rights as more absolute and will resist trading rights for benefits. In contrast, those for whom the principle of benefit maximization is central will hold that rights are justified because of their consequences and thus may be altered or waived when their consequences are undesirable. Suppose, for example, it is found that a high level of due process in teacher evaluation is inefficient in that it reduces administrative discretion that is, in turn, important to the success of school programs. This will be a more persuasive consideration for those for whom the principle of benefit maximization is central. Those for whom the principle of equal respect is central may insist that the rights of teachers be respected even at some cost to program efficiency.

Ethical Principles for Evaluation. The following ethical principles seem especially important to teacher evaluation.

Due process: Due process requires evaluative procedures that ensure that judgments are reasonable. Among its basic elements are that judgments be made on the basis of known and reasonable standards that are consistently applied from case to case, that judgments be made on the basis of reasonable evidence, and that there be systematic and reasonable procedures for collecting and testing evidence. (See *Goldberg v. Kelly*, 1970, for a judicial formulation.)

Privacy: Privacy involves a right to control information about one's self and protects people from unwarranted interference in their affairs. In evaluation, it requires that evaluative procedures not be overly intrusive and that evaluation pertain only to those aspects of a teacher's activity that are job related. It also protects the confidentiality of evaluation information.

Equality: Equality, in the context of evaluation, can best be understood as a prohibition against making decisions on irrelevant grounds Peters, 1966, chap. 3). Generally such characteristics as race, religion, gender, or ethnicity will be considered irrelevant.

Public Perspicuity: The principle of public perspicuity requires openness to the public concerning evaluative procedures, their purposes, and their results.

Humaneness: The principle of humaneness requires that consideration be shown for the feelings and sensitivities of individuals in evaluative contexts.

Client Benefit: The principle of client benefit requires that evaluative decisions be made in a way that respects the interests of students, parents, and the public in preference to those of educational institutions and their staffs.

Academic Freedom: Academic freedom requires that an atmosphere of intellectual openness be maintained in the classroom for both teachers and students. Evaluation should not be conducted in a way that chills this environment.

Respect for Autonomy: Teachers are entitled to reasonable discretion in, and to exercise reasonable judgment about, their work. Evaluations should not be conducted so as to unreasonably restrict discretion and judgment.

The ethical principles sketched above can be expressed in a more detailed set of rights that should govern the practice of teacher evaluation. These are detailed in Table 21.1.³ Note that while the general ethical principles described above should characterize any system of evaluation, the bill of rights below tends to assume a view of authority over evaluation that in Section II will be called *bureaucratic democracy*.

Table 21.1 A Bill of Rights for Teacher Evaluation

Rights of Educational Institutions

1. Educational institutions have the right to exercise supervision and to make personnel decisions intended to improve the quality of the education they provide
2. Educational institutions have the right to collect information relevant to their supervisory and evaluative roles.
3. Educational institutions have the right to act on such relevant information in the best interest of the students whom they seek to educate.
4. Educational institutions have the right to the cooperation of the teaching staff in implementing and executing a fair and effective system of evaluation.

Rights of Teachers

1. Professional rights
 - a. Teachers have a right to reasonable job security.
 - b. Teachers have a right to a reasonable degree of professional discretion in the performance of their jobs.
 - c. Teachers have a right to reasonable participation in decisions concerning both professional and employment-related aspects of their jobs.
2. Evidential rights
 - a. Teachers have the right to have decisions made on the basis of evidence.
 - b. Teachers have a right to be evaluated on relevant criteria.
 - c. Teachers have the right not to be evaluated on the basis of hearsay, rumor, or unchecked complaints.
3. Procedural rights
 - a. Teachers have the right to be evaluated according to general, public, and comprehensible standards.
 - b. Teachers have the right to notice concerning when they will be evaluated.
 - c. Teachers have the right to know the results of their evaluation.
 - d. Teachers have the right to express a reaction to the results of their evaluation in a meaningful way.

Table 21.1 continued

- e. Teachers have the right to a statement of the reasons for any action taken in their cases.
- f. Teachers have the right to appeal adverse decisions and to have their views considered by a competent and unbiased authority.
- g. Teachers have the right to orderly and timely evaluation.
4. Other humanitarian and civil rights
 - a. Teachers have a right to humane evaluation procedures.
 - b. Teachers have the right to have their evaluation kept private and confidential.
 - c. Teachers have the right to evaluation procedures that are not needlessly intrusive into their professional activities.
 - d. Teachers have the right to have their private lives considered irrelevant to their evaluation.
 - e. Teachers have the right that evaluations not be used coercively to obtain aims external to the legitimate purposes of evaluation.
 - f. Teachers have the right to nondiscriminatory criteria and procedures.
 - g. Teachers have the right not to have evaluation used to inhibit the expression of unpopular views.
 - h. Teachers have the right to an overall assessment of their performance that is frank, honest, and consistent.

Principles of Conflict Resolution

1. Remediation is to be preferred, where possible, to disciplinary action or termination.
2. Mediation is to be preferred, where possible, to more litigious forms of conflict.
3. Informal attempts to settle disputes should precede formal ones.

Rights of the Public

1. Parents and the members of the community have a right to expect that the educational welfare of children will be the paramount concern in any system of teacher evaluation.
2. Parents and the members of the community have the right to expect that their children will be taught by competent people.
3. Parents and the members of the community have the right to expect that the competence of teachers will be assessed on a regular basis and in a fair and functional way.
4. Parents and the members of the community have the right to expect that evaluation will be acted on in a way that improves the education of their children and protects their children against maliciousness or incompetence.
5. Parents and the members of the community have the right to have their concerns and complaints fairly considered.
6. Parents and the members of the community have the right to have teachers evaluated according to publicly known standards and by publicly known practices.

II. Professionalism and Democracy

In this section the author considers the question of legitimate authority over the process of evaluation. He proceeds by describing three normative models of the organization of schooling and their implications for a conception of teacher evaluation. These are called *bureaucratic democracy*, *professionalism*, and *communitarian democracy*.

These are ideal types. No institutionalization of evaluation is likely to or should exemplify any one of them fully. Moreover, in the public institutions of any society in which the commitment to the sovereignty of the people is central, *democracy*, and in large institutions, *bureaucratic democracy* will be dominant. Public schools are paradigmatic of such institutions. Thus, those who hold that the values of professionalism or of communitarian democracy should also receive expression within these institutions will face the task of showing how political values other than those of bureaucratic democracy can be served within institutions where the political values of bureaucratic democracy must be central.

The purpose of this analysis is to show that how we organize schools and how we institutionalize evaluation depend on more than what we believe about what is required for good education or what we believe about what is required for effective teacher evaluation. Our institutional forms also reflect our basic political values. This being so, proposals for changing current institutional forms must be assessed in light of these values. Further some tensions within the governance structures of educational institutions can be profitably explored by reference to the tensions between these political values. This exploration can be illuminating for how teacher evaluation is institutionalized.

Professionalism and Bureaucratic Democracy

Bureaucratic democracy has these characteristics. First, it identifies democratic control over education with legislative control. Educational policy should be made by one or more elected bodies. Such authority is constrained both by the possibility that elected officials can be voted out of office and by an independent judiciary that protects various rights. In an organization of any size, democratic authority is exercised by members whose duty is to implement legislatively determined policy. The authority exercised by such managers will be bureaucratic in its structure. Authority will be hierarchally organized. Managers will generate rules and procedures that implement policy, and they will monitor compliance.

Under such a system, while employees are subject to the supervision of managers, the authority they are subjected to is democratic authority. While judgment on various matters may be delegated to managers or employees, the chief virtue of such judgments is that they should be subservient to democratically achieved policy. In a democracy, for employees to challenge or subvert such policy outside of the normal democratic process is to deny the sovereignty of the people. It is crucial to the argument of this chapter that it be recognized that in such organizations the authority of managers or bureaucrats expresses this sovereignty.

School districts organized as bureaucratic democracies structure their evaluation systems along certain lines. Teacher are employees of school districts. School districts have certain democratically achieved objectives to accomplish

for which they hire teachers. They set up a management framework and an organization system to accomplish these objectives. One of the functions of management of school districts is to monitor teachers to ensure compliance with policy. Thus, teachers are (supposedly) judged by administrators primarily in terms of how effectively they contribute to the accomplishment of externally determined objectives.

The ethics of evaluation under bureaucratic democracy will emphasize the rights and duties that teachers and management owe one another. Teachers owe loyalty to the organization's legitimate democratically achieved purposes and to management, so long as it seeks to implement these purposes. Management owes teachers a duty of fair treatment. Evaluations must be conducted in a just and reasonable manner. Often, working out a detailed view of the ethics of evaluation will be seen as a matter of balancing a concern for fair treatment of teachers against a concern for accomplishing the school's objectives. The principles of benefit maximization and respect for persons suggest general criteria for judging when a suitable balance has been struck.

Claims by teachers to be members of a self-governing profession challenge the framework of bureaucratic democracy and, as a consequence, require us to rethink the framework concerning the nature of legitimate authority in which teacher evaluation is embedded.

Kimball provides a useful characterization of the profession:

This analysis has yielded a prevailing definition of a profession that incorporates two fundamental characteristics. First, a profession is said to be a guild, meaning that a profession constitutes a voluntary association whose membership is defined by occupation and is self-regulating and self-disciplined. Second, a profession is said to be founded upon a "science." That is to say, the occupation of the profession requires command of a theoretical body of knowledge whose acquisition involves several years of advanced study apart from an apprenticeship, usually at a university (1988, p. 3).

Consider now the following remarks by Linda Darling-Hammond concerning professionalism in education:

Professionalism is not blind conformity to state, local or even school-level policies. At the core of the definition of a profession is the notion that its members must define and enforce their own standards of practice. As Barber puts it, "An essential attribute of professional role is autonomy and self-control regarding the development and application of generalized knowledge in which they [professionals] alone are expert" (1985, p. 212).

Later she remarks:

Ultimately professionalism requires collective control by teachers over the technical decisions that define teaching work and collective responsibility for the appro-

priate treatment of students. This means not only peer review of practice, but also peer involvement in the prevention of malpractice. It also means a reconception of administration as a support function for teaching rather than a mechanism for the control of teaching (1985, p. 214).

These comments suggest⁴ that professionalism sees a connection between authority (or control) over education and the possession of a knowledge base. That the knowledge base for teaching requires judgment for its application is a reason why the work of teachers is not susceptible to direct bureaucratic control. Moreover, the knowledge base of teaching is esoteric. Only teachers are expert in this knowledge. Presumably the profession of teaching is grounded in a body of knowledge that can only be acquired by long and rigorous training and is unavailable to those who have not undergone this training. Thus the judgments that teachers can make because of their possession of this body of knowledge cannot be made by others.

The control claimed for teachers is limited. (Presumably it is relative to the kind of knowledge one believes that teachers possess.) In Darling-Hammond's view, control is to be over "the technical decisions that define teaching work and collective responsibility for the appropriate treatment of students." Thus, it does not amount to any right on the part of professional teachers to broad control over the aims of education or other types of policy, for example, tax rates. However, this authority includes control over the process of evaluation. It requires peer review of practice and peer involvement in the prevention of malpractice. Finally, the administrative role is to become a support function instead of a mechanism for controlling teachers. Thus, professionalism involves significant reconstruction of the managerial function.

This justification of teacher authority is nondemocratic in two ways. First, arguments for representative democracies appeal to the idea that people have an equal right to participate in collective decision making. Government is to be not only for the people, but of and by them. In a democratic society, the people are sovereign. Professionalism, however, sees expertise as a justification for authority. As Kimball (1988, p. 4) observes, "... the crucial factor to note about the relationship between the two points in the prevailing structural analysis of the professions is this reciprocity: the science serves to legitimate the authority of the guild, and the guild exercises authority over the social function to which the guild pertains." Thus, the fundamental commitment of professionalism is to the proposition that "Those who know should rule." As Plato (1962) noted, this commitment is an alternative to democratic views of authority. In short, professionalism is nondemocratic in that it appeals to political values other than those of popular sovereignty to legitimate its authority.

Second, because professionalism is rooted in nondemocratic values, it follows that any successful institutionalization of professionalism is also nondemocratic to the extent that professionalism actually prevails.⁵

Of course, in public institutions professionalism cannot be institutionalized in a way that simply disregards democratic sovereignty. As a practical matter public schools are going to be continue to be governed by school boards and to hire administrators to implement policy. Responsible advocates of professionalism do not suggest otherwise. But is there then not a conflict between the values of professionalism and the requirements of democratic control? Kimball characterizes this conflict as follows.

To be sure, conditions vary from state to state, as do the constitutional provisions and statutes that govern education, and in certain respects, local lay control is circumscribed by educational bureaucrats in state departments of education. But even these state departments of education are subject to politically appointed or elected officials, who are usually not educators. And, in any case, it remains true generally in this country that a person with an axe to grind or a favorite project to promote or a cousin who needs a job can run and be elected to the local school committee, thereby becoming vested with authority to determine curriculum, choose textbooks, and make educational policy. This is precisely the authority that the putatively excellent, competent, well-paid, well-trained professional would not possess (Kimball quoted in Goldstein, 1976, pp. 1355-1356).

Kimball then notes:

This relationship [between professions such as law and medicine and public agencies] is totally opposite to the one obtaining between a school board and the educational professionals under its purview. There is scarcely a "professional" decision that a teacher makes that the local politician or used-car salesman on the school board is not legally empowered to reverse. The law does vary from state to state, but as a general rule "neither sound constitutional analysis nor authoritative precedent support a federal constitutional right of teachers to determine what they teach contrary to the desires of school authorities superior to teachers within the state-sanctioned chain of command" (Kimball quoted in Goldstein, 1976, pp. 1355-1356).

Kimball goes on to claim that advocates of professionalism in education have not adequately addressed this difficulty and that professionalism in education must fail because teachers lack legal authority over those matters about which they claim expertise. Author Strike does not find either of these conclusions evident. The legal institutionalization of professional authority is being addressed. (See for example New York State Commissioner's Task Force on the Teaching Profession, 1988.) Moreover, it is not evident that the relations of authority between elected officials, their hired managers and teachers cannot be altered so as to permit space for professionalism. Instead Strike would conclude that since professionalism is inconsistent with democratic bureaucratic authority, the institutionalization of professionalism in education must be accomplished in a way that erodes that authority. That is the import of the suggestion that the role

of administration in education must be reconceptualized away from a control function and toward a service function.

The reconceptualization of authority required would have the following features. First, ultimate authority of legislative bodies and their appointees must be recognized. Second, the exercise of this authority must be constrained in such a way that operational authority for the control and evaluation of their own work is effectively vested with teachers. The center principle of this institutionalized forbearance on the part of authorities must be to administer in ways which scrupulously avoid the substitution of lay judgment for that of teachers on those substantive matters about which teachers are expert. Democratic authority must be exercised without involving lay persons or administrators in the practice of teaching. Finally, while this vesting of authority in teachers might be thought of as a delegation of authority, this sort of delegation must be distinguished from the sort that maintains the full and direct control of democratic bureaucratic authority. In any organization decisions may be delegated to subordinates. But this delegation is easily undone. School boards or administrators who allow teachers to select their own texts may simply change their minds. However, when authority is vested in professionals by legislation or contract and when it is backed by a general acceptance of the values of professionalism, it is less easily retrieved. Administrators or school boards who attempt to do so may find themselves on the wrong end of grievances, law suits or accused of abusing their authority.

In short, not only are the moral foundations of professionalism different from those of democracy, but the institutionalization of professionalism requires some erosion of democratic authority. One has not institutionalized professionalism if school boards and administrators retain the same power to control and monitor the work of teachers that they now exercise. Beating the sword of administrative control over the work of teachers into the plowshare of service to that work requires the inhibition of the exercise of democratic bureaucratic authority.

That professionalism is nondemocratic in this way is not a sufficient reason to be against professionalism, but it is telling about the sort of argument required to justify it. Consider that liberal democratic societies have a strong presumption in favor of two mechanisms of accountability.⁶ The first is free choice, consumer sovereignty. Producers of goods and providers of services are accountable to consumers who need not purchase their products. Second, when free choice does not provide effective protection for consumers and when collective decisions are required, decisions should be made democratically. Producers and providers are then accountable to democratic authority. The presumption in favor of democratic control has been quite strong concerning public education. We need not assume that this assumption is irrebuttable, but we should ask when is it reasonable, given this presumption, to vest significant authority over a practice in its practitioners. A successful argument for professionalism must show two things. First, individual choice or democratic control over practice must be insufficient to protect individuals against incompetent practice or insufficient to protect the

public interest. Second, vesting control over practice in licensed practitioners must be an effective way to prevent the harm that can result from incompetent practice and to protect the public interest.

Two claims show that a given profession satisfies these conditions. First, competent practice in the profession requires possession of an esoteric body of knowledge. Not only does this body of knowledge ground practice, but it grounds it in such a way that those who do not possess it are not readily capable of identifying competent practice or competent practitioners. Second, practitioners of the profession are subjected to a period of professional socialization that is sufficient to ensure mastery of the body of knowledge and subservience to the norms or mores of the profession. Thus, practitioners are competent and can be expected to govern their activities by the proper ends of their profession and by a concern for the welfare of their clients. Professionalism provides a means of accountability that is likely to be successful when the character of decisions puts them beyond the competence of democratic authority.

It follows that a *sine qua non* of the claim that a given practice is a profession is that it possess such a body of knowledge and that there are suitable professional norms into which practitioners can be initiated. Conversely, lack of a suitable knowledge base or ethical confusion about proper aims and norms are grounds for rejecting the claim that a given practice is a profession.

Consider what it would mean for teacher evaluation to treat teaching as a profession. Professional standards are applied in evaluation in a profession in three ways. First, professional standards are crucial for admission and initiation into the profession. Because professionals are responsible to the standards of professional practice and to their peers, and, because they are less directly accountable to any democratically constituted body or administrative authority, it is especially important to guard admission in order to be assured that aspirant professionals possess the appropriate knowledge base and that their practice is governed by it. Generally admission into a profession is accomplished by a licensure exam and by a lengthy internship. Standards of achievement are set and applied by the profession.

Second, professionals enhance their skills collegially by such activities as reading appropriate research, attending professional meetings, and discussing problems with their colleagues. Collegial relations are a vehicle to facilitate competent self-evaluation, which is the norm. It is assumed that the professional activities of professionals are not directly monitored by the school principal or other administrators or even by peers performing day to day inspection of day to day inspection of the teacher's practice. It is to be expected that professionals will exercise considerable autonomy in their activities.

Finally, although self-evaluation and collegiality are the norms under professionalism, there are times when a professional's activities have obviously fallen sufficiently short of the standards of the profession to merit sanction or exclusion. Here the professional may be accused of malpractice. In such cases where

a professional's ethics or competence become intolerable, the profession may exercise coercive authority against its members.

The ethical context of evaluation for the ideal types of bureaucratic democracy and professionalism differ. The former is dominated by the assumptions that teachers owe a duty of loyalty to externally determined goals and policies and to an organization that hires them to implement these goals and policies and that in turn the organization owes them a duty of fair treatment. In the professionalism model, teachers when they are engaging in their professional activities, owe their loyalty to clients, to professional standards and to the profession.⁷ In schools where some mix of these models of authority are institutionalized, divided loyalties are possible. Administrators who insist that teachers act contrary to their professional judgment thus generate conflicts of loyalty and ethical dilemmas for professional teachers.

Implementing these practices for teachers under the professionalism model may require significant structural changes in the organization of schools and some stratification in the teaching professional. They will require some role differentiation among teachers. Novices may have to be distinguished from licensed teachers, and those more senior teachers on whom the burden of monitoring novices or evaluating other teachers falls may have to be distinguished from other licensed teachers. Mechanisms should also be developed whereby collegiality becomes an important factor in the professional development of teachers. Also, mechanisms for identifying and dealing with malpractice are required. This would presumably involve some codification of professional and ethical standards as well as means for their application. Finally, in the professionalism model, the administrative role is not the bureaucratic "bridge of authority" between legislative authority and teachers whereby the work of teaching is controlled.

These arguments suggest three important conclusions. First, the professionalization of teaching requires a significant erosion of bureaucratic democratic authority over teaching practice. Second, professionalism is likely to result in significant restructuring of the management of educational organizations. Finally, the argument that teaching is a profession depends very substantially on the claim that there is an esoteric body of knowledge in which the practice of teaching is grounded and that there are suitable professional norms for the regulation of practice.

There are other reasons for considering the professionalization of teaching. Treating teachers as professionals may be a means of enhancing the status, income, and working conditions of teachers. This is required if we are to recruit competent teachers in sufficient numbers. Also, it is claimed that one of the ills of schools is an excess of bureaucratic organization (Wise, 1979). The bureaucratic domination of teachers not only demeans teachers and degrades working conditions, it denies teachers the autonomy of judgment they need to be effective, and it undermines the organizational conditions for effective planning and

program development (Chubb, 1988; Grant, 1988). Professionalization may be seen as a cure for these ills.

We need not reject these claims. But absent a demonstration that teaching is grounded in an esoteric body of knowledge, we need not treat these arguments as arguments for professionalization. We may see them as reasons for rolling back the scope of bureaucratic regulation of schooling, for providing a higher level of autonomy at the building level or for encouraging broad teacher participation in school governance. Teacher empowerment need not be synonymous with professionalization. Proposals to professionalize teaching are more than suggestions that we may improve education if we treat teachers better or that we may have better teachers if we educate them better. They are also proposals to substitute nondemocratic institutional forms for democratic forms. They are proposals to significantly alter our political values about the governance of education. Thus, arguments for the professionalization of teaching must be sufficiently powerful to override the presumption in favor of the bureaucratic democratic control of teaching practice. Such arguments must demonstrate the incompetence of democratic governance to make the judgments on which competent teaching and competent teacher evaluation depend.

If this is correct, it is crucial to ask whether a suitable knowledge base for teaching and suitable professional norms exist or might soon exist.⁶ In the absence of a suitable knowledge base, the institutional changes required by the professionalization of teaching would amount to a usurpation of democratic authority, one that restricts the democratic bureaucratic monitoring and control of teachers work and substitutes for it a fictional alternative of a profession regulated by nonexistent standards.

The author cannot here attempt to decide the extent to which there is such a knowledge base. (For various candidates, see Brophy and Good, 1986; Gage, 1986; Darling-Hammond, 1985; and Shulman, 1987.) The best case for treating teaching as a profession relies on the expertise of teachers in subject matter. It seems reasonable, for example, to suggest that it should be beyond the power of school boards and administrators to require teachers to lie about the laws of physics, misrepresent the facts of biology, or to treat poetry as though it was solely a means for gaining employment with greeting card companies. Moreover, it seems clear that it is beyond the competence of lay persons to decide if instruction in many academic disciplines is faithful to the subject matter. Few administrators will be able to decide if the content of instruction in physics or calculus is accurate. Evaluation of subject matter competence where the subject matter is an academic discipline requires expertise in that discipline. To the extent that pedagogical practice is dependent on subject matter expertise that too cannot be evaluated by those who lack subject matter competence. However, it is not clear that this argument is successful in gaining professional status for those teacher who do not teach disciplines. Moreover, it tends to conceptualize teachers as members of various academic professions rather than as professional teachers. Nor is it clear that there is a body of pedagogical knowledge sufficient

to ground teaching as a profession. It may be plausibly argued that there is little in educational scholarship that is both esoteric and noncontroversial and that many teachers practice quite effectively without possession of the putative knowledge base. Likewise, confusion about the ethics of teaching or standards of malpractice (*Donohue v. Copiague*, 1979) count against teaching being a profession.

Thus, the claim that a suitable knowledge base exists or will soon exist is somewhat short of self-evident.⁹ Moreover, and more importantly for the argument of this chapter, there are other political values that might inform how we govern education and how we institutionalize evaluation that should be considered. For these reasons, the author wishes to explore one additional model of educational authority, rooted in a different set of political values, and its implications for evaluation.

Communitarian Democracy

Consider three different sets of values that might be considered important to a full view of democracy.

Democracy as fair representation: One view of democracy holds that the crucial thing about democracy is that everyone's interests in collective decision making should be represented fairly. It is crucial to respect both the legitimacy of everyone's interests and the equal right of everyone to representation in decision making. Typically, such sentiments are expressed by institutions that attempt to represent everyone through a fair electoral process, together with attempts to restrict undue influence by special interest groups on the legislative process. Congress, state legislatures, and school boards are attempts to express these political values.

Democracy as nonhierarchical decision making: This concept of democracy is concerned that no one should be dominated in the process of decision making. Domination is understood as a relationship where one person has the power to command another without that person's participation in the decision-making process. Undominated relations will be those in which each is a direct and equal participant in the decision-making process and an owner of the decision.

Democracy as dialogical decision making: This conception of democracy emphasizes that decisions should be reached by a process of rational discussion or dialogue. Notions of free expression are especially important to this vision of democratic decision making.

These sets of values that are important to democracy are not mutually exclusive. Indeed, they contain a common core. Each values freedom, equality and rationality. Nevertheless, they express different emphases about what is crucial in democratic decision making that may, in some circumstances, prove competitive.

Most obvious is the potential tension between democracy as representation and the second and third values of democracy. In a representative democracy, two things will be crucial. First, each individual in the relevant community must have fair representation in the process whereby the legislature is constituted. Second, the decisions of the legislative body must be respected. Fair electoral processes are required to satisfy the first concern. The second concern is satisfied by procedures that ensure that the will of the legislature is faithfully carried out by those who implement it. This is the view of democracy that in large organizations is expressed in bureaucratic democracy.

The second and third values of democracy, by contrast, suggest that we should be reluctant to see people located in hierarchical and bureaucratic organizations and that we should prefer that decisions be made in organizational structures where those who must implement or undergo group decision are full participants in the discussions that produce them. A view of school governance that emphasizes these values of democracy is far more likely to see schools as democratic communities in which decisions are achieved by dialogue between those individuals who are within the walls on a regular basis. (See Gutmann, 1987; Raywid, 1988.)

The virtues of this communitarian view of democracy are that it allows for high levels of participation in decision making by those most directly affected, it makes decisions dialogically, and it tends not to create hierarchical structures. Relations of dominance are thus avoided. The liability is that it seems to exclude or reduce the role of the external political community in decision making.

It is possible to argue for teacher empowerment by appealing to the ideals of communitarian democracy. Nonhierarchical and dialogical decision making applied to schools will legitimate teacher participation in school governance. Moreover, a concept of educational governance articulated in the name of communitarian democracy will have advantages over professionalism.¹⁰ It will legitimate teacher participation in decisions without delegitimizing other members of the community. Parents, other members of the community, and students will not be excluded from participation because they are not regarded as suitably knowledgeable.

Communitarian democracy requires the construction of a functioning school community, a community in which decisions are achieved by discussion among the administrators, teachers, and families (parents and students) of a particular school. Applied to evaluation, it requires a higher level of dialogue among the parties involved in the education of children. It enfranchises teachers and their immediate clients (parents and students) as legitimate participants in the process.

The suggestion that parents and students should be involved directly in teacher evaluation will surely be resisted by some. Yet they are involved. Students resist inept teachers and become discipline problems. Parents complain, lobby, and try to rearrange their children's schedules to avoid those whom they regard as poor teachers. Such happenings inevitably influence administrative judgment. Why not seek organized ways to include parents and students in actual

evaluative decisions? Doing so would empower them, would allow them to contribute their views in a legitimate way, might enhance their sense of ownership of decisions, and would subject their influence to an organized process and to the requirements of due process.

Communitarian democracy requires decentralization of educational decision making and a considerable diminishment of bureaucratic control of teaching. It makes teachers primarily responsible, not to a bureaucratic hierarchy or to a set of professional standards and a peer group (although elements of each would remain), but to the most immediate community in which they work and which they serve. In such schools, teacher evaluation might be thought of as a community function in which each of the parties contributes to the discussion that which they are in the best position to contribute. For example, families might offer their assessment of the effect of teaching on the learning of children, on their morale, and on how teaching serves their educational aspirations. Teachers will contrib-

Table 21.2 Models of Legitimate Authority in Evaluation

<i>Bureaucratic Democracy</i>	<i>Professionalism</i>	<i>Communitarian Democracy</i>
<i>Primary Locus of Control</i>		
political community	peer judgment autonomous practice governed by internalized standards of practice	school community (families, teachers, students and administrators)
<i>Central Justificatory Principle</i>		
equal representation of interests in democratic decision making	expert decision making	non-hierarchical and dialogical decision making
<i>Means of Control</i>		
policy making by elected body bureaucratic implementation to ensure compliance with legislative will	peer-controlled entry collegiality shared norms of practice peer-monitored code of ethics	community decisions shared, agreed upon values
<i>Evaluation</i>		
democratically achieved policy	peer-controlled licensure	community participation
management monitoring of employed compliance	apprenticeship monitoring	discussion of issues and problems
teachers owe duty of loyalty to policy goals	self-evaluation: loyalty to clients and professional standards	contributions according to capacity, training and degree and manner of involvement
administrators owe duty of fairness to teachers	peer judgments of malpractice	

ute subject matter competence as well as pedagogical wisdom. Administrators might contribute a knowledge of the more technical features of evaluation and represent the interests of the larger political community beyond the schools.

Table 21.2 shows a comparison of the features of the ideal types of bureaucratic democracy, professionalism, and communitarian democracy as both political value systems and as applied to the institutionalization of evaluation. Currently bureaucratic democracy seems to dominate educational governance and institutional forms. Arguably, bureaucratic democracy has evolved so as to degrade the work of teachers and disregard their judgment and so as to exclude the local community from direct participation in decision making. Professionalism is a response to the first of these problems. However, it does little to address the second and might even worsen it. Communitarian democracy, by contrast, suggests an outlook which is less bureaucratic in its forms and which also empowers both teachers and the local community. It, like professionalism, tries to keep the political community at arms length. Also like professionalism, it will value the wisdom of teachers in those areas where they are knowledgeable. Unlike professionalism, it will not see expertise as a justification for enhanced authority.

Any view of educational governance and of the institutional forms that will structure evaluative practice must respect the central political values of each model. The real trick is to integrate and balance them in responsible ways. As currently the values of bureaucratic democracy seem to be dominate, perhaps we need to move in ways that respect teachers *and* that include the local school community in decision making. Reform movements should be judged, not only on their consequences for teachers, but on the extent to which they create educational communities.

Summary

In Section I of this article the author reviewed a number of ethical principles that must be taken into account in any system of teacher evaluation. This account was concluded with a Bill of Rights for Teacher Evaluation that applies most directly under a bureaucratic democratic structure of evaluation. Section II deals with issues of legitimate authority in evaluation. Here are reviewed three models of educational authority and their implications for teacher evaluation.

Notes

1. For a more extensive development of these principles and an application to a variety of educational problems including those of evaluation, see Strike and Soltis, 1985; and Strike, Haller, and Soltis (1988).
2. For discussion in the context of school finance, see Strike (1988).
3. This is adapted from Strike and Bull (1981). The argument for these rights is made there.

4. The comments that follow are an interpretation of what I believe is involved in a reasonable conception of professionalism. While I use the remarks of Darling-Hammond to introduce the discussion, my interpretation goes beyond her remarks. No claim that she would agree with what follows is intended.

5. This point follows logically from the previous one. That is, if in fact the justification of professionalism is nondemocratic, the claim that a successful institutionalization of professionalism is nondemocratic is a tautology. Thus, if this argument is to be attacked, it must be attacked on the first claim, not the second.

6. This formulation is indebted to Bull, 1988.

7. An important issue that I cannot address here is the role of unions in the professionalization of teaching. Traditionally unions assume that the relationship between teachers and schools is that between management and labor. Unions thus pursue teachers' rights against management. It is not clear that unions can simultaneously perform as professional associations. There may be role conflicts or conflicts of interest. For example, the same organization should not function to protect teachers against unfair evaluations and simultaneously have power over substantive evaluative decisions.

8. I propose four criteria for such a knowledge base.

(1) It must be esoteric. That is, it must be possessed by virtue of long and rigorous training. Knowledge that could be possessed by many cannot be the basis of a profession since then everyone could judge competent performance. Other forms of accountability could succeed.

(2) It must be the basis of the characteristic judgments and activities of the profession. Educational scholarship sports numerous esoteric branches of knowledge, educational philosophy and educational psychology, to name two. It is not, however, clear that the practice of teaching is rooted in such enterprises. The knowledge base of teaching must ground practice.

(3) It must be the basis of evaluation in the field. It must ground the process of admission and initiation into the profession, be the basis of collegial discourse, and support judgments of malpractice.

(4) It must show how school teaching differs from other forms of teaching in our society. The idea that we might professionalize all teaching in our society and restrict the practice of teaching to those who have been professionally trained is absurd and would violate basic rights. It would require, for example, that we license parents. We must then know why it is important for school teachers to possess this esoteric body of knowledge while it is not similarly important for the others who teach in our society.

Because teaching is an emerging profession, we need not insist that these standards be met immediately. But we should at least insist that the advocates of teaching as a profession show that there are reasonable prospects for meeting them.

9. There are three things that do not follow from a denial that there is a suitable knowledge base to ground teaching as a profession. First, it does not follow that some teachers are not more competent than others or that we cannot discover this. Not all competence is grounded in an esoteric knowledge base. Second, it does not follow that nothing of value is known or worth transmitting about how to teach. There is considerable distance between knowing something worthwhile and possessing an esoteric body of knowledge. Finally, it does not follow that teachers are ignorant or stupid. Discussions about the knowledge base have to do with the state of the art, not the competence or capacity of its practitioners.

10. A pure communitarian democracy also has some liabilities. Perhaps it is plausible to believe that individual school communities should have more authority over educational programs, and school boards and state legislatures less. But imagine school communities claiming authority over tax rates. We should soon be reminded that "no taxation without representation" is a central commitment of our society.

References

- Bentham, Jeremy. (1961). An introduction to the principles of morals and legislation. In J. Bentham & J. S. Mill, *The utilitarian*, (pp. 7-398). Garden City, NY: Doubleday.
- Brophy, J. J. & T. Good. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching*, (3rd ed.) (pp. 328-375). New York: Macmillan.
- Bull, B. (1988). *Professionalism and the nature of teacher autonomy*. Unpublished manuscript.
- Chubb, J. E. (1988). Why the current wave of school reform will fail, *The Public Interest*, 90, 28-49.
- Darling-Hammond, L. (1985). Valuing teachers: The making of a profession, *Teachers College Record*, 87(2), 205-218.
- Donohue v. Copiague Union Free School District*, Court of Appeals of New York, 47 N.Y. 2d 440 (1979).
- Gage, N. O. (1986). *Hard gains in the soft sciences: The case of pedagogy*. Bloomington, IN: Phi Delta Kappa.
- Goldberg v. Kelly*, 397 U.S. 254 (1970).
- Goldstein, J. G. (1976). The asserted constitutional right of public school teachers to determine what they teach. *University of Pennsylvania Law Review*, 124(6), 1293-1357.
- Grant, G. (1988). *The world we created at Hamilton high*. Cambridge, MA: Harvard University Press.
- Gutmann, A. (1987). *Democratic education*. Princeton, NJ: Princeton University Press.
- Kimball, B. A. (1988). The problem of teachers' authority in light of the structural analysis of professions. *Educational Theory*, 38(1), 1-9.
- Mill, J. S. (1961). The utilitarians. In J. Bentham & J. S. Mill, *The utilitarians*. (p. 399) Garden City, NJ: Doubleday.
- New York State Commissioner's Task Force on the Teaching Profession. (1988). *The report on the commissioner's task force on the teaching profession*. Albany, NY: Author.
- Peters, R. S. (1966). *Ethics and education*. Glenview, IL: Scott, Foresman.
- Plato. (1962). *The republic*. New York: Oxford University Press.
- Raywid, Mary Anne. (1988). Community and schools: A prolegomenon. *Teachers College Record*, 90(2), 197-210.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-22.
- Strike, K. A. (1988). The ethics of resource allocation: Questions of democracy and justice. In D. H. Monk & J. Underwood (Eds.), *Miscellaneous school finance: Issues and implications for policy* (pp. 143-180). Cambridge, MA: Ballinger Publishing Co.
- Strike, K. A. & Bull, B. (1981). Fairness and the legal context of teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 303-343). Beverly Hills, CA: Sage Publications.
- Strike, K. A., Haller, E. J. & Soltis, J. F. (1988). *The ethics of school administration*. New York: Teachers College Press.
- Strike, K.A. & Soltis, J. F. (1985). *The ethics of teaching*. New York: Teachers College Press.
- Wise, A. E. (1979). *Legislated learning*. Berkeley, CA: University of California Press.

22. Governance Issues in the Evaluation of Elementary and Secondary Schoolteachers

Arthur E. Wise

Tamar Gendler

Earlier chapters have reviewed the purposes of teacher evaluation and the methods appropriate to each. This chapter considers some of the governance issues associated with each of the eight major purposes of teacher evaluation. For each purpose, it describes traditional relationships and current developments and then considers the implications of emerging teacher professionalism. *Governance* refers to the exercise of authority and influence by government agencies, quasigovernmental agencies, and nongovernmental agencies. The primary foci of concern are shifts in governance among levels (local, state, national agencies) and among agencies at each level.

Teacher professionalism refers to the adoption by schools and teachers of control mechanisms employed in the established professions. Thus, teacher professionalism seeks to promote:

- Reform of teacher education so that prospective teachers are well educated and knowledgeable about teaching. Increasingly, this preparation is seen as requiring a four-year liberal-arts education, a fifth year of graduate-level professional and pedagogical preparation, and a year-long supervised internship, preferably in a professional development school.
- Reform of teacher licensing so that prospective teachers must demonstrate that they have acquired the knowledge and developed the skills necessary to practice independently. Increasingly, licensing is seen as requiring the creation of state professional practices boards to establish and enforce meaningful standards for entry to teaching.
- Reform of schools so that they can accommodate the professional practice of teaching. Increasingly, this change is seen as requiring teacher participation in school planning and decision making so that educators articulate, share, practice, and enforce a common set of educational standards.
- Reform of teacher organizations to balance their union and professional functions. Increasingly, this reform is seen as requiring the National Education Association

AUTHORS' NOTE: We appreciate the advice of our reviewers Charles Kerchner and Gary Sykes. We have heeded most of their advice.

(NEA) and American Federation of Teachers (AFT) to collaborate with states and school districts to establish workable procedures for identifying incompetent teachers and remedying their deficiencies or removing those who cannot overcome them.

- Reform of testing practices so that student assessment encourages good educational results and professional teaching practices. Increasingly, these practices are seen as requiring essay examinations and the evaluation of student projects and products with teachers involved in developing outcome criteria and grading beyond their own classrooms.

Teacher evaluation referring to the assessment of teaching performance. Thus, the chapter will consider the implications of professional control mechanisms for evaluating teaching performance with respect to the eight identified purposes of teacher evaluation: preservice; selection; certification and licensure; assistance to and assessment of beginning teachers; tenure and retention; career advancement and merit pay; and school improvement.

Preservice Evaluation

Historically, responsibility for preservice evaluation belonged exclusively to schools of education. Free to decide whom to admit, to continue and to graduate from their programs, the schools relied mostly upon traditional academic measurements for their decisions. Admission was based on successful completion of the freshman or sophomore course of study, or occasionally upon achievement of a minimum grade point average. Continuation in the program rested on attaining passing grades in courses required or offered. Graduation was contingent upon successfully meeting school of education (and other university) course requirements. Most courses were academic, with grading based a student's performance on tests, papers, and related academic exercises. The major exception was student teaching, where the student's grade was based on actual classroom performance.

Such a system assessed little more than the prospective teacher's capacity to perform as a student in an academic setting. Students who were strong academically and who were sophisticated test-takers would do well. Conversely, those who were academically weak or poor test-takers would do poorly. How tough the grading was varied from campus to campus and from professor to professor. The tacit hypothesis, in education as in other fields, was that the better a person performed as a student, the better he would perform as a practitioner.

The one exception to this was student teaching, which was generally viewed as the place where a prospective teacher could reveal talent for teaching. However, the ratio of supervising teachers to student teachers was low. As a result, supervising teachers, because they lacked intimate familiarity with the performance of student teachers, often gave them the benefit of the doubt — and a passing grade.

In the 1970s, plummeting enrollment in teacher-education programs put schools of education onto the horns of a dilemma: either significantly curtail programs — including staff and budgets — or keep standards for entry and graduation low. As more and more schools chose the latter option, states began to intervene, mandating minimum scores on standardized tests and minimum grade point averages for admission to teacher education (Darling-Hammond & Berry, 1988, pp. 11-14).

The effect of these requirements has been to transfer responsibility for admissions decisions from the school of education to the state and, in the case of standardized testing, to change the criteria by which prospective teachers are selected. Because such tests are given very early in the student's college career, they may eliminate from teacher education those who have inadequate education in high school but might come to flourish in the college environment. Of course, neither the standardized test nor the minimum GPA is a measure of teaching ability. Thus, a legislative decision does eliminate early from teaching some who might learn how to be teachers by attending a school of education.

From a governance perspective, the decision-making discretion of a unit of the university has been severely curtailed. What was once the province of the university has been limited by legislative fiat. The mandate makes the decision; schools of education and professors are less able to make judgments about whom to admit. The faculty senate at the University of Wisconsin noticed the precedent-shattering nature of the decision (Darling-Hammond & Berry, 1988). It noted that if the legislature could set standards for entry to teacher education, it could similarly set standards for entry to all fields. Moreover, if the legislature could create standards in time of surplus, it could again eliminate them in times of shortage.

Teacher Selection

Teacher selection is and must remain the responsibility of school districts. The school district, under the existing structure of school governance, is officially the employer of teachers. Technically speaking, the school board has the hiring authority which it delegates to the administrative staff at the central office and school site. The actual responsibility is then in the hands of school administrators, a few of whom involve teachers in the selection process. Especially in large districts, teacher involvement is not the norm.

The traditional criteria, methods, and processes and their outcomes are reasonably well understood in large districts (Wise, Darling-Hammond, & Berry, 1987). A district's operational definition of the good teacher depends on the tools used, the values and influence of decision makers, and the nature of the information used to screen candidates. Because school districts usually commit few resources to the selection process, administrators tend to use low-cost (e.g., transcripts, references, etc.) to medium-cost (e.g., tests and structured inter-

views, etc.) information. Generally, they do not use high-cost approaches such as on-the-job assessments or demonstration teaching.

In addition to lack of resources, administrators lament logistical problems, including the need to handle large numbers of candidates evenhandedly and the lack of coordination and timing between recruiting and screening, which inhibit the use of data on teaching performance. Thus, some districts turn to other types of information. One medium-cost approach is the commercially developed, standardized screening interview instrument, which is relatively inexpensive to administer, allows comparison across candidates, and provides a quantitative ranking of prospective teachers.

However, no publicly available scientific evidence supports the validity of virtually any of the instruments in use. Some instruments may, in fact, eliminate well-qualified candidates. In one case the author examined, some of the questions posed and the scoring procedures used systematically discriminate against candidates who are cognitively oriented. For example, an interviewer asks two candidates, "Why should we hire you?" The first candidate answers that she can make the most difficult concepts easy to understand, and the second one answers that he loves children. The scoring procedures require the interviewer to rate the former response negatively and the latter response positively.

Second, our investigation revealed "antiintellectualism" among some school district decision makers. The research literature confirms that the most academically able teachers are not necessarily the most likely to be hired. While some districts do pay considerable attention to the academic qualifications of teacher candidates, other administrators report that they prefer teachers with average academic ability because they have seen "too many straight-A students who didn't know anything" about teaching. Some administrators believe that the most academically able teachers do not possess the ability to work with diverse students and parent groups, the desire to work in extracurricular assignments (especially athletics), and attitudes that mesh with local community expectations.

Third, even when school districts attempt to uphold high academic standards in the screening process, the needs of a particular vacancy may actually determine who gets hired. For example, a principal who needs a football coach and a mathematics teacher may hire a less-qualified mathematics teacher who can also coach. Or a principal who needs an English teacher and a sponsor for the junior-senior prom may hire a less-qualified English teacher who is willing to sponsor the prom. Extracurricular demands and limited staff allocations are realities for principals and schools.

The process and the logistics of teacher hiring may have at least as much influence on the quality of the staff hired as do formal screening mechanisms. First, a school district's discourteous treatment of candidates may indicate to some prospective teachers that they are not valued. Teachers report considerable difficulty in obtaining vacancy information from district personnel offices, many of which are understaffed. When applicants are able to "get through," it is

generally to a secretary who may not be able to respond to their queries. Papers are lost; vital information is often not conveyed. Prospective teachers may wait months for any indication of whether or not they will be hired.

Second, a school district's management information system must satisfy the logistical demands of reporting vacancies, analyzing data, and communicating the data to decision makers. However, inaccurate projections of teacher vacancies, delays associated with district and other governmental budget decisions, late resignations of teachers, and most significantly, the lengthy process of emptying the internal transfer pool can create this unfortunate scenario: a school district has found an excellent candidate but cannot hire him or her immediately because there is a question of whether or not there will be a vacancy to be filled. Often when the question is answered, the candidate is no longer available.

In addition, in large school systems with thousands of applications, an antiquated system of "piles of files" often makes it difficult to match applicants' skills with job specifications efficiently and effectively. Not all school decision makers have easy and equal access to the information and therefore are forced to choose the first candidate who seems good enough.

Reform of teacher selection at the district level is necessary, though, strictly speaking, reform of governance is not (Wise, Darling-Hammond, & Berry, 1987). In particular, teacher involvement may be a level for altering the criteria, methods, processes, and outcomes of teacher selection. Teacher involvement can enhance the validity of the process by providing greater insight into candidates' subject matter competence and teaching philosophy.

Certification and Licensure

The traditional use of the term *certification* in teaching varies from the common use of the term in the established professions. Certification in teaching is the award by the state of a certificate to practice. This governmental decision in professions is generally called *licensing*. The state grants a license to practice to professionals. Certification in professions refers to a determination by a professional body that a novice is fit to practice. In the professions, professional certification and governmental licenses have generally been merged as states rely for the licensing decision on procedures devised by professional bodies.

In the licensed professions, the state has delegated to each profession the responsibility to determine who shall receive a state license. Generally, the state legislature establishes a professional licensing body made up primarily of members of the profession. These bodies often include some faculty members drawn from the appropriate professional school and some include lay members. The actual members of the professional licensing boards are appointed by governors, often in consultation with the legislature. Once appointed, these boards establish the qualifications for entry to the profession and determine who may enter.

These state licensing boards often combine to form national federations that agree to common standards. Over the years, these federations have had nationalizing influences. While only the state has the legal authority to establish and enforce standards, the state boards come to embrace the common standards and use the uniform procedures for licensing developed by the national federations. However, conflicts periodically erupt as one or another state takes exception to federation policy.

Currently, teacher licensing (usually called *certification*), in all but a handful of states is directly controlled by a government agency, the state department of education. This difference, among many others, prevents teaching from being "a profession" as one of the main defining characteristics of a profession is that it has responsibility over entry to its ranks.

The effect of state governance is significant. Until recently, state department governance has coincided with a certification process tied to graduation from an approved program of teacher education. If a candidate successfully completed an approved program of teacher education, he or she was automatically eligible for certification.

Approval of teacher education programs has also, in many states, been in the hands of the state department of education. This, too, distinguishes teaching from professions where separate nongovernmental accrediting agencies, made up of the professional schools themselves, determine which professional schools will be accredited.

In recent years, state departments of education have begun to establish requirements for licensing beyond graduation from an approved program. States have begun to require tests of basic skills, subject-matter knowledge, and professional and pedagogical knowledge. In addition, they may now require assessments of teacher performance (Darling-Hammond & Berry, 1988). In addition, states are paying increasing attention to the National Counsel for the Accreditation of Teacher Education (NCATE).

These changes *in procedure* are a major step toward bringing teacher licensing into line with the professions. They establish procedures and requirements independent of the professional school and its requirements for graduation. Professions have all found it necessary to disentangle responsibility over licensing from responsibility over professional preparation.

The rationale is clear. The profession and the public need to be assured by an independent procedure that a novice is fit to practice. The independence and integrity of that process is essential. The professional licensing body must be free from political influence. It must especially be free to establish and enforce standards, independent of pressures generated by trends in demand and supply.

A handful of states, including Minnesota and Oregon, have established professional standards boards to develop and enforce standards for teaching. Proposals to establish new boards or to strengthen existing advisory boards are being advanced in a number of states.

Those who advocate standards boards believe that vesting the responsibility for licensing in a state agency with multiple missions threatens the integrity of the licensing process. State departments rightfully are concerned not only with teacher quality, but also with teacher supply. When shortages of teachers develop, state departments fall under pressure to reduce or eliminate standards for entry to teaching, thus compromising the integrity of the licensing process.

Standards boards advocates believe that for standards to be maintain, responsibility must be in the hands of a board which is insulated from political pressure. Moreover, state departments, while they have been creating new requirements for teaching, have not shown a propensity to wrestle with the thorny intellectual questions about professional standards. Instead, they have placed too much reliance on inexpensive paper-and-pencil, multiple-choice tests which do not reveal a candidate's teaching knowledge or skills. The solution in the professions, all of which have had to wrestle with criteria, methods, and procedures, has been to establish boards which have the responsibility to establish and enforce standards at the state level.

Newly on the horizon in teaching is the National Board for Professional Teaching Standards, most of whose members are teaching professionals. This body, established with the assistance of the Carnegie Corporation, appears to have decided to focus its attention upon advanced certification for teachers. Having made this decision, the Board, for the present, has removed itself from concerns over initial teaching licensing, though the Board has influenced the national climate with respect to teaching standards and teacher testing. At this juncture, it is not clear whether the Board will play the role played by other national professional bodies.

Assistance to and Assessment of Beginning Teachers

Assistance to and assessment of the performance of beginning teachers is a recent development in teaching. This development is motivated by two somewhat different considerations. One is the widespread recognition that schools of education cannot allocate sufficient time and resources to the development of teaching skills. Thus, beginning teachers, upon hiring, should be given assistance in developing their skills. The second is the well-founded belief that paper-and-pencil tests cannot adequately determine who should receive a teaching license. Hence, beginning teachers should be assessed during their first year of teaching. Nearly two dozen states have adopted the procedure (AACTE, 1986). The procedure has been required by legislative mandate or state board or department of education decree. In most instances, though, the burden for staffing the assistance and assessment functions has fallen on local district personnel, with little, if any, state financial support. As one result, the preponderance of effort has been directed at the assessment function. This propensity is

reinforced by the fact that the assessment function is easier to enforce than the assistance function. Typically, there is a state beginning teacher evaluation instrument that must be completed.

The assistance function is likely to receive short shrift if it relies upon existing resources to get the job done (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). Over the years, some school systems have operated "buddy" systems wherein new teachers are assigned a buddy, another teacher in the school. And, certainly, many senior teachers voluntarily offer assistance to junior colleagues. However, in these circumstances, neither the senior teacher nor the new teacher has time during the school day to work together. Nor do they easily have time to converse in the demanding schedule that is the worklife of most teachers.

Effective teacher assistance requires that senior teachers have their time at least partially reallocated so that they can provide assistance to new teachers. Moreover, their job descriptions and assigned responsibilities need to reflect this shift. In other words, they must officially be expected to be mentors to new teachers. It is also possible that beginning teachers should be given less than full responsibility for a less than full schedule.

Because effective teacher assistance is labor intensive, it is expensive. It is not likely to occur if states expect the cost to be borne by local districts. The responsibility of local districts is the education of students and, while a district may benefit in the long run from an investment in assistance to its beginning teachers, it may also not benefit. Teachers, especially many who begin in urban or rural districts, often transfer from them after the first few years of teaching. Thus, an urban district may not reap the benefits of an investment in beginning teachers. Because the assistance function may be construed as part of the training of new teachers, it is appropriate for the state to assume financial responsibility.

The proper allocation of costs is tricky. Interns in professions are generally paid because they provide services. However, in medicine, for example, teaching hospitals receive funding from a variety of sources including federal and state governments which recognize the responsibility of the teaching hospital for the training of interns.

Thus, as a first step, the state must provide funds to local districts which are willing to provide assistance to beginning teachers (Wise, Darling-Hammond, with Berry & Klein, 1987). Only if the financial burden is accepted by the state will assistance programs operate effectively.

Concerning assessment, states must clarify purposes. Assessment may be carried out to make a licensing decision or a continuation of employment decision. The latter is the clear responsibility of school districts; the former is the responsibility of the state. If the purpose of assessing beginning teachers is to decide whether or not they should continue to be employed by the school district, then the responsibility is and should remain that of the school district.

However, much of the impetus for assessing beginning teachers has its origin in state-level dissatisfaction with paper-and-pencil tests as the means to determine a new teacher's competence. If the purpose of assessment is to help

determine whether a new teacher should receive a full license, then control over the assessment must be in the hands of the state. For an assessment process to be appropriate for a state licensing decision, it must be uniform and meet standards of reliability and validity. The process cannot be left under the control of numerous school districts which will inevitably have different standards and be subject to different degrees of employment pressure. Letting school districts decide who should receive a state license is rather like allowing law firms to decide who should be a lawyer. School districts in several southeastern states are now required to assess their beginning teachers to help the state decide whether they should be given a license. Variable standards could mean that some new teachers will be denied the opportunity to teach not only in the district but also (undeservedly) in the state. At the same time, other teachers could not only be allowed to teach in a district but also (undeservedly) be granted a state license.

Affixing responsibility for assessing performance at the state level, preferably in the state professional practices board, is the necessary first step. The standards for teaching and the procedures for assessing the standards must be statewide. The granting of a license by the state is the recognition that an individual is at least minimally qualified to teach anywhere in the state so far as the state licensing authority is concerned. Practically speaking, however, states, depending upon the procedure they employ, may have to rely on personnel not in the state's employ. Assessing the performance of teachers requires the time of senior personnel. It must be clear that when these personnel are assessing teachers for the state, they are acting for the state and under its direction. Current checklist evaluation procedures make relatively modest time demands. However, as performance assessment becomes more sophisticated, assessments will require more time and the costs will rise. These may be borne by the state as part of the licensing process or the state could require local districts to bear the costs. Of course, the state may also choose to pass the costs on to the candidate as it does in the professions where candidates pay substantial fees to be subjected to the licensing process.

Evaluation for Tenure and Retention

Tenure is a status granted after a trial period to teachers protecting them from dismissal without due process. Tenure, generally as a matter of state law, creates a presumption of continued employment within a school district. Tenure laws define causes and procedures for the dismissal of tenured teachers. Tenure, then, is a privileged employment status which is granted by a school district decision, acting under state law.

Typically, a teacher will be granted tenure status after a short period of trial employment, often two or three years. Theoretically, the nontenured teacher is to be subjected to close supervision and scrutiny during the trial period. Upon successful evaluation, tenure status is awarded. Theoretically, the procedures for

dismissal of nontenured teachers should be different from the procedures for dismissal of tenured teachers. Often, however, school districts employ the same evaluation instruments and procedures for the evaluation of beginning teachers, continuing teachers, and teachers who are candidates for dismissal (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984).

While school districts evaluate teachers for tenure and retention, they would do well to recognize the presumptions evident in law. First, school districts should carefully scrutinize beginning (pretenure) teachers. Where states establish beginning teacher programs, the local district's activities are likely to be prescribed. Where successful completion of the first year of teaching is a condition for a state license, the state is likely to prescribe criteria, methods, and procedures.

Under policies now being conceived, a hiatus remains between the successful completion of the first year and the explicit or implicit tenure decision. The decision is often implicit because it is made by virtue of a decision to grant another year or two of employment. School districts may wish to consider establishing special evaluation procedures for this hiatus period as traditional, bureaucratic evaluation procedures do not uniformly prevent the granting of tenure to those who should not receive it. Thus, local decisions implementing state tenure laws result in the grant of a special employment status in a particular district. Unlike the state's decision to grant a license, tenure has no meaning beyond the school district in which it is granted despite the fact that state law is involved.

Because the granting of tenure creates the presumption of continued employment, the school district should design evaluation procedures for tenured teachers consistent with the idea most teachers need evaluation solely for improvement. Districts should design formative evaluation procedures which serve this function.

School districts must also have exceptional evaluation procedures for use in instances where a tenured teachers is judged to be a potential candidate for dismissal. *Exceptional* connotes *management-by-exception* as it is used in the management literature. Work (in this instance, teaching) is allowed to continue without direct managerial intervention, unless an *exception* is noted. Where this occurs, an exceptional procedure is invoked. In teaching, such procedures have been found to work where they are collaboratively designed and implemented by management and labor (Wise, et al., 1984). A teacher who is separated from a school district is free to seek employment in other districts. Indeed, under state law, the teacher can teach until and unless the teaching license is revoked by state authority.

Evaluation for Teacher Improvement and Professional Development

Evaluation for teacher improvement and professional development is a school district and individual responsibility. As an individual responsibility, it implies the acceptance of a professional ethic of continuous development — the need to improve oneself and to keep abreast of new knowledge.

As a school district responsibility, the link between teacher improvement and professional development is obviously determined by the nature of the district's evaluation system. Evaluation systems designed to screen for minimal competence or for retention or promotion are unlikely to be compatible with the demands of improvement. Evaluation systems may, however, be tailored to the requirements for improvement. In particular, evaluation systems which encourage teachers to develop their own goals for improvement may help teachers to improve (Wise et al., 1984).

The terminology is murky, however. *Teacher improvement* appears generally to be about improvement based upon systematic feedback from district officials. Most districts do not operate differentiated teacher evaluation systems linked to performance improvement. While virtually all teacher evaluation systems contain the expectation that teachers will receive feedback, the nature of the feedback is generally limited. In part, these limits are caused by the widespread use of one-size-fits-all evaluation models. In part, they are caused by the practice of using generic evaluators, administrators without specific subject-matter or grade-level expertise.

The problem is not a governance issue; school districts have the right and responsibility to ensure competent performance by their employees. Rather, the problem is the nature of standard practice. School districts seriously interested in the use of evaluation to improve performance need to design systems which utilize expertise. As teachers gain in experience, they can receive more constructive advice from those who are experts in the relevant teaching fields and methods.

It goes without saying that improvements in performance need not be linked to an evaluation system as formally defined in educational practice. Teachers can, introspect, read, take a course or be given advice by colleagues, students or administrators as a basis for improvement. In this sense, *teacher improvement* is closer to the term *professional development*. An individual teacher's professional development is his or her own responsibility.

The shift of the burden from the district to the individual is a major shift in governance. If a teacher is responsible for his or her professional development, then he or she must decide what development is necessary. The individual must assume the burden of evaluating his or her needs and must select experiences which will fill those needs. Professional development for teachers, like the analogous function in other professions, is a professionally imposed and self-administrated procedure. The profession, sometimes with the sanction of the state,

may impose requirements for continuous or periodic professional development. Typically, the requirement is stated in terms of numbers of courses, credits or hours. Professionals then incur an ethical or legal obligation to take courses, if not to improve. The individual professional must meet the requirement, but, of course, is free to engage in other developmental activities.

It is common in professions for professional associations to facilitate the continuous development of their members. They offer courses, seminars, and other learning experiences directly. They also work with professional schools, individual professors, and other agencies to arrange experiences.

To the extent that professional development is a professional responsibility, the role of evaluation (except for self-evaluation) recedes and the nature of the learning experience itself becomes paramount.

Evaluation for Career Advancement/Merit Pay

Career-advancement, career-ladder, or merit-pay programs may be district, state, or national in scope. Districts may design and operate career-ladder programs as in Charlotte-Mecklenberg. Teachers may be evaluated for promotion in a career ladder that creates new ranks or new positions or for considerations for additional pay within the district. Presumably, any special such considerations would not convey to the teacher if he or she leaves the district. States may design and operate programs as in Tennessee. Presumably, new ranks or positions attained by a teacher would have transferability across district lines within the state. Career-advancement programs may operate nationally as may occur under the auspices of the National Board for Professional Teaching Standards. Recognition under such a program would attach to the teacher wherever he or she went, though whether rank, position or salary attached to that recognition would be a local or state determination.

Career ladders for teachers may simply recognize increasing levels of competence as is the case for college professors. College faculty continue to have the same job as they are promoted through the professorial ranks. The promotion from assistant to associate professor presumably recognizes growing competence on the job rather than being an indication of readiness for a new job. In such a system, the criteria for evaluation are criteria related to competence or performance on the job. In the case of college teachers, the competence evaluated involves teaching and research. In the case of elementary and secondary teachers, the competence is teaching. The purpose of the recognition determines the criteria to be used. Such programs can operate within a district, within a state, or even nationally (by consent). Persons who transfer across district or state lines could be recognized by rank. Such recognition is largely the case with reference to education — masters degrees are almost universally rewarded by school

districts. In the same way, recognition as a Level I, II, or III teacher could be accomplished.

Most career ladders have, however, operated on an additional premise. While recognizing increasing levels of competence in the job of classroom teachers, they contain the expectation that higher ranked teachers will have additional or different responsibilities. Thus, to the extent that competence in teaching is not coordinated with competence in other educational activities (mentoring, curriculum development), evidence must cover a broader spectrum of activities. Career ladder programs of this type are most logically operated at the district level since presumably only the district knows how much teacher leadership it can accommodate. Programs at the state level will inevitably be unable to match teaching talent to district needs. Shortage of highly ranked teachers in some districts and surpluses in others are the inevitable result. Even local programs must struggle to balance the aspirations of teachers with the slots for leadership available. In any case, the state would have to evaluate not only teaching but also extra classroom leadership functions. Criteria, methods, and processes would be designed and operated so that appropriate emphasis is given to both.

At the state and national level, it is probably easier (but by no means easy) to recognize increasing levels of teaching competence. By comparison, advanced certification programs in the professions that operate them recognize competence in the function of the profession not in the capacity to exercise leadership in the profession. In medicine and accounting, advanced certification implies that the individual has high levels of knowledge and skill. The evaluation processes are designed to determine competence to do the job.

To be sure, a district or state may well require advanced certification as a condition for providing leadership. A school district or a state may insist that a teacher be board-certified as one of the requirements for teacher leadership. However, the local district would be well advised to operate additional evaluation processes to determine qualifications for teacher leadership positions.

There are no widely accepted (written) tests for leadership. Schools, in selecting teachers for administrative posts, have used classical criteria and procedures familiar to schools, industry, and government. Candidates for leadership must undergo scrutiny of their teaching competence (which may or may not rank high on the decision criteria), advanced education (MA, ED.D.), extraclassroom leadership capacity, etc. The classical method is to review past performance by examining formal evaluations, and interviewing colleagues and superiors and to judge prospective performance by interviewing the candidate about his or her views.

Evaluation for career advancement is thus best done with an eye toward purpose. The criteria, methods, and processes should especially reflect the distinction between recognizing teaching competence and identifying candidates for teacher leadership. The former can be done at the local, state or national level; the latter requires processes at the operating level.

Teacher Evaluation for School Improvement

The use of individual formal teacher evaluation results as a guide to school improvement or collective staff development is not common. School improvement and staff development in America's schools are generally motivated by developments external to school districts. While staff development may be directed at improving teachers' skills, the programs are at best only loosely coupled to an analysis of the collective needs of teachers as revealed by evaluation results.

The absence of a linkage in this area points out one of the serious consequences of the implicit bureaucratic management model. The notion that the evaluation of teachers should be reflected in school improvement and staff development is based on the idea that managers will evaluate individual teachers and then plan collective activities to remedy their deficiencies. Yet, school managers have insufficient expertise and time to evaluate teachers and to plan and implement remedial activities.

Another vision is that teachers collectively and collaborately with managers assess their needs and planning for staff development and school improvement (Darling-Hammond, 1986). This shift, while not representing a shift among levels of government, does restructure the functions at the school or school district level. Teachers and administrators jointly assess the state of practice in the district and plan staff development and school improvement activities. In this way, the expertise of the teaching staff is unleashed and brought to bear on evaluating practice, planning and delivering staff development, and improving school performance.

Conclusion

The recent educational reform movement has intensified interest in teacher evaluation for all purposes. The assessment of teaching performance is becoming more pervasive. If teacher evaluation is to move from a perfunctory bureaucratic activity carried out in response to external pressures to improve teachers and teaching, then the methods and the control of teacher evaluation must change. If teacher evaluation continues to be construed as an activity done by administrators to teachers, it is not likely to improve either teachers or teaching. Improvement in the methods of evaluation alone will not be sufficient. There will never be enough administrators or sufficient pedagogical and subject-matter expertise among administrators to meet the growing demand for teacher evaluation. If states and school districts continue the classical hierarchical model of teacher evaluation, they will perpetuate the simplistic, generic approaches now in use. If teacher evaluation is to play an important role in improving teachers and teaching, then control over the process must be shared. At the state level, this requires a structure — a professional standards board — to allow professional

teachers to articulate and enforce performance standards. At the local level, it requires that teachers share the responsibility for quality control over their colleagues. These changes will demand new mindsets on the part of teachers and their unions and on the part of school administrators and school boards at the state and local level. Much will depend on the ultimate influence of the National Board for Professional Teaching Standards and the extent to which its model of shared responsibility becomes the norm at the state and local level.

References

- American Association of Colleges for Teacher Education (AACTE). (1986, Dec.) *Teacher education in the states: A 50-state survey of legislative and administrative actions*, Washington, DC.
- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, 35, 281-322.
- Boulton, D. L. (1973). *Selection and evaluation of teachers*. Berkeley, CA: McCutchan.
- Bridges, E. M. (1986). *The incompetent teacher*. Philadelphia: Falmer Press.
- Broudy, H. S. (1956). Craft or profession? *The Educational Forum*, (January), 175-184.
- Burks, M. P. (1986). *Requirements for certification for elementary schools, secondary schools, junior colleges 1986-87*, (51st ed.) Chicago: University of Chicago Press.
- Darling-Hammond, L., & Berry, B. (1988, March). *The evolution of teacher policy*. Santa Monica, CA: The RAND Corporation. JRE-01.
- Darling-Hammond, L., Gendler, T. & Wise, A. (1989). *Toward a professional assessment system for licensing teachers: An interim report to the Minnesota board of teaching*. Santa Monica, CA: The RAND Corporation.
- Gage, N. L. (1978). *The scientific basis of the art of teaching*. New York: Teachers College Press.
- Gendler, T. (1989). Staking a path to professionalism, *Basic Education*, (January), 8-10.
- Horwitz, R. A. (1979). Effects of the 'open classroom.' In H. J. Walberg (Ed.), *Educational environments and effects: Evaluation, policy, and productivity*. Berkeley, CA: McCutchan.
- Levine, M., & Gendler, T. (1988). Background paper: Professional practice schools. In M. Levine (Ed.), *Professional practice schools: Building a model*. (pp. 27-70). Washington, DC: American Federation of Teachers.
- Mastain, R. K., (Ed.) (1984, 1988). *The NASDTEC manual: Manual on certification and preparation of educational personnel in the United States*. Sacramento, CA: National Association of State Directors of Teacher Education and Certification.
- McKeachie, W. J., & Kulik, J. A. (1975). Effective college teaching. In F. N. Kerlinger (Ed.), *Review of Research in Education*, Vol. 30. Itasca, IL: F. E. Peacock.
- Mickler, M. L., & Soloman, G. L. Beyond credentials in teacher selection: A validation study of the Omaha teacher interview. *North Central Association Quarterly*, 59, 339-345.
- Niece, P. (1983). The interview and personnel selection: Is the process reliable and valid? *The Clearing House*, 56, 232-255.
- Peterson, P. L. (1979). Direct instruction reconsidered. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching*. Berkeley, CA: McCutchan.
- Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, 35, 1-62.
- Rudner, L. M., Project Director. (1987, Aug.) *What's happening in teacher testing: An analysis of state teacher testing practices*. Washington, DC: U.S. Government Printing Office.
- Shavelson, R., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research*, 51(4), 455-498.
- Shive, J. J. (1988, Nov. Dec.) Professional practices boards for teachers. *Journal of Teacher Education*, 39(6), 2-7.
- Smith, J. M. (1980). Teacher selection procedures, *The Clearing House*, 53, 312-314.
- Tyson-Bernstein, H., Wise, A. E. & Darling-Hammond, L. (draft, 1987, May, forthcoming). *Testing teachers: The content and control of entrance examinations*. Santa Monica, CA: The RAND Corporation. WD-3397-CSTP.

- Wise, A. E. (1979). *Legislated learning*. Berkeley: University of California Press.
- Wise, A. E. (1986). The case for trusting teachers to regulate their profession (commentary). *Education Week*, 6, 24.
- Wise, A. E. (1987, Oct.) *Center for the study of the teaching profession: Annual report: October 1986 to September 1987*. Santa Monica, CA: The RAND Corporation.
- Wise, A. E. (1988, Nov. 16). If we are ever to 'professionalize' schoolteaching, universities must redesign education programs (Opinion). *Chronicle of Higher Education*.
- Wise, A. E. (1989, Jan. 11). States must create standards boards (commentary). *Education Week*.
- Wise, A. E., & Darling-Hammond, L. (1984, Dec. 1985, Jan.) Teacher evaluation and teacher professionalism. *Educational leadership*, 28-31.
- Wise, A. E., Darling-Hammond, L., McLaughlin, M. W. & Bernstein, H. T. (1984, June). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: The RAND Corporation. R-3139-NIE.
- Wise, A. E., Darling-Hammond, L. & Berry, B. (1987, Jan). *Effective teacher selection: From recruitment to retention*. Santa Monica, CA: The RAND Corporation. R-3462-NIE/CSTP.
- Arthur E. Wise, Darling-Hammond, L. (with Barnett Berry & Stephen P. Klein). (1987, Nov). *Licensing teachers: design for a teaching profession*. Santa Monica, CA, The RAND Corporation. R-3576-CSTP.
- Yoder, W. H. (1976). A new approach in teacher selection. *Illinois School Research*, 12, 21.
- Young, I. P. (1984). The effects of interpersonal performance style in simulated teacher selection interviews. *Journal of Research and Development in Education*, 17, 43-51.
- Young, I. P., & Allison, B. (1982). Effects of candidate age and teaching experience of school superintendents and principals in selecting teachers. *Planning and Change*, 13, 245-256.

23. Economic Aspects of Teacher Evaluation

Stephen A. Hoenack

David H. Monk

I. Introduction

Interest in teacher evaluation is a byproduct of concern with the effectiveness of schools in achieving desired learning outcomes. To the extent that teachers' effectiveness can be improved, and especially if weaker teachers can either be dismissed or strengthened, evaluation can potentially raise the instructional productivity in terms of enhanced learning for students and increased rates of return to education.

There are alternative methods of evaluation and each has many variants. They can be costly and their effects on teachers and indirectly on students can be complex and not entirely beneficial. Consequently, any proposed evaluation process should itself be evaluated according to its benefits and costs in relation to those of alternatives, including the option of not taking any evaluative action.

Economics provides a framework that can assist in identifying variables to be considered and it can serve as a helpful guide for organizing the information gathered when analyzing evaluation alternatives. The purpose of this chapter is to explore what economics has to offer those with interests in teacher evaluation. It begins with a discussion of selected areas of economics that have bearing on the design, implementation, and ultimately the evaluation of teacher evaluation. Next, attention turns to alternative forms of teacher evaluation and related economic issues. The goal is to demonstrate linkages between the areas of economics discussed previously and the practical design, implementation, and the evaluation questions raised by specific forms of teacher evaluation.

AUTHOR'S NOTE: We wish to acknowledge with thanks the assistance provided by the two consultants for this chapter: Professor Byron W. Brown of Michigan State University and Dr. Roy Dexheimer, District Superintendent of Schools, Tompkins, Seneca, and Tioga Board of Cooperative Education Services, Ithaca, New York.

Finally, we examine some of the implications for the administration of teacher evaluation from a macro perspective beyond the level of the individual institution.

II. Relevant Areas of Economics

Several aspects of economic analysis can help improve understanding of the benefits and costs of evaluation. The discussion will deal separately with the following topics relevant to evaluation: production theory and efficiency; the economics of information; incentives; and the distributional effects of policy interventions.

Production Theory and Efficiency

The central idea of production theory, as it has been applied to the study of the learning process, is that the contributions of different educational inputs toward achieving learning outcomes can be estimated and compared. This information coupled with knowledge of the inputs' relative costs has relevance for efforts to improve efficiency where the goal is either to minimize the total cost of reaching a given learning outcome or, alternatively, to maximize the level of outcome for a given total cost. A closely related idea concerns the substitutability of inputs where economists presume it is often possible to produce a given learning outcome using alternative combinations of inputs. For example, a particular learning outcome might be achievable by replacing some of the time and effort of teachers with more intensive use of material resources such as books and computers.

The concepts of production theory and of economic efficiency are as applicable to teacher evaluation as to learning. When, for example, skilled classroom observers are costly, it may be possible to economize by substituting less expensive video recordings of teaching for live classroom observations. Over a longer time, the high price of classroom observers may induce technological improvements, such as the use of video-disk technology to simulate different teaching situations and evaluate teacher responses.

The efficient choice of evaluation resources can depend on the scale of production. Some resources, such as videotaping and criterion referenced tests may have high development costs but low variable application costs. Hence they would more likely be desirable choices when there is a large number of evaluated situations.

Evaluation can involve differing relative emphasis on the outputs and inputs of the educational production process. An output oriented evaluation has the advantage that evaluators do not need to know as much about the educational production process as in an input oriented evaluation in order to make correct inferences about teacher productivity. However, there are problems, to be dis-

cussed in Section III, if all of the important educational outcomes are not measured. Moreover, appropriate interpretation of the information obtained in an output oriented evaluation can depend on the teacher's and community's weights given to different learning outcomes.

A useful input oriented evaluation requires that the evaluator as well as the teacher understand the production process. This process may differ from situation to situation and may depend on attributes such as the characteristics of students, the subject matter being conveyed, the scale of activity, and the overall budget for inputs. That is, alternative situations can affect the relative cost-effectiveness of the educational inputs employed. If evaluators are not aware of these contingencies, they can make incorrect judgments about teachers' productivities.

The Economics of Information

Evaluation is the task of acquiring and interpreting information about educational productivity. The analysis of choices in developing and utilizing information is a new and rapidly developing field within economics. Evaluation may be regarded as an application of this subfield, often referred to as the economics of information.

A key concept is that while better decisions often can be made with more complete information, the costs of added information should be justified by the resulting incremental value of desired outcomes. The costly information obtained in a teacher evaluation about school activities or learning outcomes is presumably for the purpose of intervening in the educational process to improve educational outcomes. A central question is, does the incremental value of this change justify the cost of the evaluation information?

An important insight offered by the economics of information is that those on the job may face lower costs than others of learning about how to be more productive. Teachers can not only select among known educational methods but decide to undergo the costs of learning about and implementing new ones. Can educational productivity be more economically raised via this knowledge obtained from practice than via research performed by specialists about possible new methods?

This question raised by the economics of information brings attention to the significance of incentives. When teachers find it personally costly to experiment systematically with methods to identify and use those that produce better learning outcomes, it may be that they should be rewarded for doing so.

Performance Incentives

A performance incentive is any aspect of the work environment that ties workers' actions to their personal interests. Costly information is usually required to establish incentives that raise efficiency. However, incentives do not have to be deliberately established, and people respond to the actual incentives

facing them, not those that should be in place. Incentives do not have to be monetary. For example, the disapproval or esteem of colleagues and superiors can act as an incentive.

Incentives have two relationships to evaluation. First, any evaluation system has incentives, whether or not they are intended. Second, planned incentives can be regarded as an alternative to traditional methods of evaluation.

A distinction has been drawn between *overall value* and *specific* incentives (Hoernack, 1983). An overall value incentive system holds an individual or group of teachers accountable for the total value of outcomes in relation to that of all resources used. In contrast, a special incentive pertains to only one aspect of the outcomes produced or only particular uses of resources while ignoring outcomes. When all important outcomes are rewarded, overall incentives are preferable to specific incentives because teachers gain from efficiently using resources under their discretion to achieve the outcomes.

Whenever teachers perceive that evaluation will directly or indirectly affect their incomes or the nonmonetary aspects of work important to them, incentives are created. If attention is not given to the incentives implicit in an evaluation system, specific incentives can be created that lead teachers to select inefficient resource uses or fail to emphasize the most important learning outcomes. For example, if merit pay is based on classroom observation by supervisors who incorrectly believe that certain teacher actions are important for learning, instructors will tend to respond by emphasizing these actions even at the expense of what is most important for students' learning.

The design of overall value incentive systems can provide an alternative to evaluations involving third parties in obtaining and interpreting costly information and then communicating it back to teachers. An example is mentioned in Section III.

The Distributional Effects of Policy Interventions

Whether intentionally or not, evaluative information can have differential impacts on the welfare of particular groups involved in the educational process. Consider, for example, Murnane and Cohen's (1986) observation that willingness to evaluate teachers more formally could satisfy taxpayers' desires for increased accountability and thereby facilitate higher salaries for teachers. To the degree that such an evaluation system fails to discriminate against less than meritorious teachers and also fails to elicit greater effort from able teachers or attract more talented teachers to the profession, there would simply be a transfer of welfare from taxpayers to teachers.

As another example, a teacher-evaluation system could work to the advantage of one type of student and to the disadvantage of another, particularly if it created incentives to change the relative amounts of teacher effort devoted to different categories of students. For example, rewards for rises in students' average measured learning gain could increase inequality when the same effort produces

more learning for relatively able students. One consequence of such an evaluation system could be an increase in dropout rates, something most decision makers would view as a negative outcome.

It is usually not practical to use some of the efficiency gains resulting from an evaluation to compensate those made worse off by it. Further, one cannot expect to find widely acceptable rules for dealing with the distributional impacts of evaluation. This is because there is no generally applicable social welfare function having non-dictatorial and other desirable properties that can balance the competing interests of students, teachers, and taxpayers (Arrow, 1963); known political mechanisms do not reflect the application of such a function. As a minimum, it is desirable for evaluators to be sensitive to and explicitly communicate impacts of alternative evaluations on the competing interests involved.

III. Applications to Alternative Forms of Teacher Evaluation

The authors distinguish four types of evaluation based on the type of information used. They are: (a) records of previous training and employment; (b) outcomes of specially designed teacher tests; (c) observations of teaching performance; and (d) measurements of student learning gains. Their economic aspects will be discussed separately, even though these types of information can of course be combined with one another.

Previous Training and Employment

Information about a teacher's formal training and prior record of service has played a large role in teacher evaluation, especially at the time of hire. Because information on educational background is economical to obtain and verify, it would be helpful to know this variable's relationship to teacher productivity. Many empirical analyses of educational production functions have expressed students' test scores as a function of measures of the teacher's formal training such as highest degree earned and years of experience, along with other variables including the student's socioeconomic background (for a recent review, see Hanushek, 1986.) Consistent relationships have not been found between the measured characteristics of teachers and students' test scores, suggesting that salary scales rewarding additional credit hours of education and years of experience may not have a close relationship to the teacher's productivity.

However, it is possible that more specific aspects of a teacher's training and experience are related to students' learning. For example, particular teacher-training courses and types of experience as a student teacher, and the nature of previous classes taught could provide useful insights into a teacher's assessment. However, in sharp contrast to the gross and readily available teacher credentials used in existing production function studies and frequently employed as incen-

tives, information that more precisely reflects the teacher's capabilities can be voluminous and costly to obtain and interpret.

Training and employment records also reflect past events and this time-bound character itself can be problematic. The meaning of a given credential can change over time (a 1975 BA may be quite different from a 1985 BA, and a 1975 practicum may be quite different from a 1985 practicum); moreover, the passage of time since the indicated event may be relevant.

This point is pertinent to new developments in credentials for teachers, many of which are emerging out of efforts to further professionalize teaching. To the extent that these credentials document teachers' performance during previous employment, their validity will need to be checked periodically. Past excellence as a teacher may be closely related to past, and perhaps difficult to maintain, contextual characteristics.

Teacher Tests

Recently, teacher tests have been the subject of benefit-cost analyses (Catterall, 1988; Shepard & Kreitzer, 1987a,b; Solmon & Fagnano, 1988.) These studies draw on the progress that has been made toward applying economic principles to practical evaluation problems (Haller, 1974; Hawley, 1987; Levin, 1983; Levin, Glass, & Meister, 1984). For example, sensitivity to the sometimes subtle nature of both costs and benefits is reflected in the Shepard and Kreitzer (1987a,b) cost analysis of the Texas teacher recertification test which looked beyond the expenditures a state makes to develop and administer a test.¹ These analysts were particularly sensitive to teachers' time to prepare for the test and the resulting foregone opportunities for other activities, including additional instruction. According to Shepard and Kreitzer's estimates, these less obvious costs constituted 93.5% of the total \$78,022,000 cost estimate.

Work remains to be done. For example, the long term costs and benefits of evaluation need to be included; the Texas test, suitably revised, may be reusable and future applications may be less or even more costly than the initial one. Another issue was pointed out by Solmon and Fagnano (1988) in connection with the Shepard and Kreitzer study. They argued that the relevant outcome measure is not the savings of dismissed teachers' salaries, but rather the subsequent differential labor market experiences of students taught by teachers who can and cannot pass the test. With a series of hypothetical assumptions linking exposure to teachers who were unable to pass the test to adverse labor market experiences (e.g., delayed job entry, unemployment, and lowered earnings because of skills not learned), Solmon and Fagnano calculated a benefit (well over 1 billion dollars) which compared favorably with Shepard and Kreitzer's estimate of costs (78 million dollars). Alternative assumptions about the relationships between exposure to non-test-passing teachers and subsequent labor market experiences would, of course, produce varying results.

Teacher tests should not necessarily be dismissed as an evaluation tool because the validity of past instruments has been questioned by some. Substantial improvements in tests may be possible, and if so, the benefits of achieving them could possibly justify the costs. A useful distinction can be made between the costs of developing and applying new tests (or any other evaluation technique.) The development problem is one of design and pretesting to make improvements that might permit more accurate inferences about the most significant skills, interests, and capabilities that contribute to effective teaching. Development costs are fixed when they do not vary with the number of instructors evaluated. In contrast, application costs may have relatively small fixed costs and vary nearly proportionately with the number of evaluated teachers. Large fixed development costs thus are easier to justify the larger the number of teaching situations to which the evaluation technique is applied.

The incentive effects of teacher tests include those to learn the material covered in them, whether or not it relates to good teaching, and to spend time and energy worrying and talking about the tests. If teacher tests were regularly used, unions and teacher groups would negotiate not only over the compensation of teachers' time studying for and taking tests but over the tests' content.

Teacher Observations

Knowledge of the educational process is a necessary condition for the development and implementation of widely applicable and reliable observation instruments and rating scales. It will be costly and, as noted earlier, risky to achieve such understanding. Application costs of teacher observations additionally depend on the degree to which routine teaching can be used for observation.

At present, we lack the necessary informational basis for a consensus about the ingredients of good teaching. The contingencies associated with production of educational outcomes seriously limit evaluators' abilities to make generalizations about teacher competence when focusing on inputs.² They also raise administrative challenges such as the need to match teaching strengths with teaching assignments. Short of a perfect match, there can arise a further need for remedial assistance to teachers found deficient in important areas. Additional resources would be required for such assistance and decisions needed about priorities in dealing with deficiencies. These decisions, in turn, require information about the costs of producing improvement in a particular area of a given teacher's range of skills.

Previous production function studies have been disappointing to those searching for a basis of such generalization (again, see Hanushek (1986) for a review), and yet some are optimistic about developing a knowledge base of what works in instructional settings (see, for example, the review by Brophy and Good [1986] as well as the claims made by Shulman and Sykes [1986]).

Application costs are based in part on the samples taken of teaching, which if small may provide inaccurate information about a teacher's actual day-to-day

contribution. This is not just a problem of random variation. Rosenholtz's (1987) study of career-ladder reforms revealed teachers voicing concern over the ability of some colleagues to be impressive only during formal assessments. The costs of all evaluation techniques are increased by the need to observe enough of the teacher's activities to make the results reasonably independent of attempts to cheat or temporarily appear more productive than is actually the case.

Several of the new teacher-evaluation proposals include a heavy reliance on observation outside of normal teaching and as a result carry some alarming cost implications. Consider the Shulman and Sykes (1986) vision of a several day assessment complete with simulated classes that are videotaped for "blind" review by experienced teachers elsewhere. The authors did not attempt a cost analysis; yet such an analysis would invite realism when thinking about unconventional possibilities.

For example, it may be possible to achieve some of the precision and standardization of a simulated class while retaining the economy of home classes by employing localized assessment centers. There is also a role to be played by technology to the degree that classes can be observed and evaluated using computing and video capabilities.

The design of observations should take into account their incentive effects. Teachers will carefully note what is looked for in the assessments and try to provide it, whether or not these attributes are productive in terms of important learning outcomes. This represents an opportunity as well as a danger since demonstrations of desired teacher behaviors within the context of evaluation can be effective means of eliciting them.

Measurements of Student Learning

The use of measured student learning gains can potentially free evaluation from the ambiguous task of inferring productivity from teaching inputs. However these data can be biased toward what is easily measured. Two challenges in employing this evaluation technique are to measure all of the important changes in learning, and to assess the degree to which they are actually due to the teacher's influence. Measured differences in the student's knowledge over the time period he or she is with the instructor tend to overstate the teacher's contribution, as other factors including the student's maturation, family influences, and prior learning experiences, can influence learning accomplishment. Offsettingly, the teacher can affect future learning.

Students' characteristics can, of course, influence their learning gains. Regression equations may have some usefulness in controlling for those influences; although there are serious difficulties with the estimation of educational production functions, it is more straightforward to control for the effects of student characteristics than for the influence of teachers and other school variables on learning.

The potential usefulness of measured learning gains is not foreclosed by the small independent effects of the teacher found in production function studies; as noted this result may be due to lack of variables controlling for effort, enthusiasm, personality, and the willingness to experiment with different teaching techniques. However, more complete inferences can be drawn from tests of students' learning only if additional resources are extended.

It is often suggested (see for example, Berk, 1984) that relatively economical standardized written tests are not as effective for inferring students' acquisition of abstract reasoning or problem-solving skills as for determining the factual knowledge they have attained. More costly testing involving oral examinations and observing students in the classroom situation may be required to infer some learning accomplishments. The one time development costs for such tests and observations are likely to be significant, and their ongoing per student application costs can be expected to be substantially higher than those of written tests. However, these costs may be well justified if they make it possible for measured learning to be a highly effective evaluation tool.

Incentive issues underline the importance of measuring all important learning outcomes. Test driven transaction is regarded as a negative outcome in large part because teachers may tend to emphasize what is tested and underemphasize what is not, even if it is important. Similarly, if in an input oriented evaluation the focus is on certain inputs such as uses of worksheets, scheduling of activities, or particular expected amounts of working individually with students, teachers may emphasize such actions even when they do not believe them to be important. A particularly undesirable incentive would be if teachers were penalized for the errors that inevitably result from experimentation with different curricula, teaching techniques, and sequencing of material.

Overall value incentives, noted in Section II, can be useful in avoiding the costs and possible perverse specific incentives resulting from input oriented evaluations. The possibility of a system that combines overall value incentives with outcome based instruction is explored in Hoenack (1988). In this system, teacher groups are jointly accountable and responsible for setting rewards for individuals. These groups are rewarded by increasing the residual gain from achieving valued learning outcomes by making efficient uses of resources.

The most important attribute of incentives based on residual gains between rewards and costs is the encouragement it can provide to experiment with alternative instructional techniques and methods of organization. The lack of such incentives explains probably more than anything else the historical lack of productivity growth in schools. The initial implementation of overall value incentives could include providing resources to teacher groups to underwrite some of the experimentation costs.

IV. Administrative Implications at a Macro Level

Previous sections dealt with detailed issues of evaluation that typically occur at the institutional level and did not consider the implications of this process at the system, state, or national levels. Here we focus on some of the macro issues of evaluation.

Trades Between Teacher and Taxpayer Welfare

Improvements in teacher evaluation can under some conditions make schools more expensive while temporarily failing to make them more productive. Suppose the improved ability to evaluate teachers is coupled with a program designed to reward teachers found to be meritorious with significant salary (or other benefit) increments, but there are no reductions in the salaries of other teachers. A meritorious teacher finally has his or her merit recognized and in this sense there is an equity gain, but unless the teacher responds by becoming even more productive the schooling system will find itself paying more for something it was already receiving for less.

However, under overall value incentives, existing productive teachers can find it in their interests to become even more valuable. In addition, over longer periods of time the evaluation system could attract new teachers of higher average ability than that of current teachers. The benefits of an evaluation system depend on the sizes of these responses to it.

Disproportionate Emphasis on Teacher Inputs

Efficiency may also be lost if a disproportionate amount of effort is devoted to the evaluation of human in contrast to material inputs. Just as it is important to evaluate teachers, it is important to evaluate curriculum, textbooks, facilities, and computer software. If reward and incentive systems are instructed so that even more emphasis is placed on undeniably costly human inputs, there is a danger of contributing further to the already heavy emphasis on teachers. A more balanced evaluation program could facilitate the substitution of inputs away from costly teacher resources and this could further enhance efficiency (Monk, 1989). The criticisms that are common of instructional software (Carnoy, Daley, & Loop, 1987) are indicative of the inadequate reward structure for nonteacher inputs. Overall value incentives giving residual gains to teachers rewards them for cost effective substitutions of these resources.

*Matching Teacher Capabilities
with Local Needs*

It is one thing to have knowledge regarding the capabilities of a given teacher; it is quite another matter to realize fully the benefits these capabilities offer. We can perhaps look forward to the day where any given teacher's range of talents and capabilities will be known with considerable accuracy and confidence. Such knowledge may even be available at reasonable cost. School districts can then be imagined to advertise for the mix of characteristics they want and to restrict hiring to those individuals possessing them. The preferred characteristics would be bid for and the market for teachers would become more efficient.

As desirable as this sounds, there is a potentially serious equity problem. The ability of administrators to match teaching talent with teaching opportunities is likely to vary, perhaps widely, across schooling settings. Poorer or smaller systems likely would face the most serious difficulties. The wealthier the district, the more successful the administrator is likely to be in bidding for teachers with the desired mix of capabilities. The larger the setting, the more flexibility there is likely to be in specifying the desired mix.

This equity issue arises with particular poignancy in cases where the improved teacher evaluation system can be used to identify teachers whose deficiencies are so widespread and serious as to make dismissal the only reasonable course of administrative action. Aside from the difficulties of actually dismissing a teacher whose incompetence is no longer in question (Bridges, 1986), there are important questions about the economic capacity of poor school districts or those in highly limited labor markets to compete with other districts for teachers having characteristics demonstrably superior to those of a dismissed teacher. Successful use of the dismissal strategy is also likely to vary across subject areas (Murnane, Singer, & Willett, 1988; Rumberger, 1987).

An irony of recent efforts to improve teacher evaluation for poor or small schools is that resources that could have been used to hire desired teachers are being used instead to provide more accurate information about the capabilities of teachers, including those these schools wish they could hire.³ This is not to say, however, that improved teacher evaluation necessarily works to the disadvantage of poor or small schooling settings. In fact, improvements in teacher evaluation may make clearer existing differences in education quality across schools and may thereby contribute to larger reform agendas.

Conclusion

Just as economic analysis can generate insight into the conduct of teacher evaluation, it can help policy makers anticipate these larger implications of proceeding with reform. We are impressed with how broadly applicable economic principles and analyses are to phenomenon such as teacher evaluation.

Our goal in this chapter has been to demonstrate some of these contributions and to highlight some areas of relevant research.

Notes

1. Unfortunately, such progress is not to be found everywhere. In the U.S. Department of Education's publication, *What's Happening in Teacher Testing* (1987, p. 6) costs to the state are treated exclusively in terms of out-of-pocket costs.
2. Examples of such contingencies include instances where the effectiveness of a teacher attribute or instructional method varies substantially across types of students.
3. For more on the administrative problems surrounding the matching of teacher talent and teaching opportunities particularly in conjunction with the effects of low secondary school enrollment, see Monk (in press, chapter 13).

References

- Arrow, K. (1963). *Social choice and individual values* (2nd ed.). New Haven, CT: Yale University Press.
- Berk, R.A. (1984). *The use of student achievement test scores as criteria for allocation of teacher merit pay*. Paper presented at the National Conference on Merit Pay for Teachers, Sarasota, Florida, March 22-23.
- Bridges, E.M. (1986). *The incompetent teacher: The challenge and the response*. Philadelphia, PA: Taylor & Francis, Inc.
- Brophy, J.E. & Good, T.L. (1986). Teacher behavior and student achievement. In Wittrock, M. (Ed.) *Handbook of research on teaching* (pp. 328-375). New York: MacMillan.
- Carnoy, M., Daley, H. & Loop, L. (1987). *Education and computers: Vision and reality*. Stanford, CA: Center for Educational Research, Stanford University.
- Catterall, J.S. (1988, April). *Estimating the costs and benefits of large scale assessments: Lessons from the past research*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Haller, E.J. (1974). Cost analysis for educational program evaluation. In James Popham (Ed.) *Evaluation in education: Current applications*. Berkeley, CA: McCutchan.
- Hanushek, E.A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 34(3), (September), 1141-77.
- Hawley, W.D. (1987). The high costs and doubtful efficacy of extended teacher-preparation programs: An invitation to more basic reforms, *American Journal of Education*, 95(2), 275-298.
- Hoenack, S.A. (1983). *Economic behavior within organizations*. New York: Cambridge University Press.
- Hoenack, S.A. (1988) Incentives, outcome-based instruction, and school efficiency. In Monk, D.H. & Underwood, J. (Eds.) *Micro-level school finance: Issues and implications for policy*. Cambridge, MA: Ballinger Publishing Company.
- Levin, H.M. (1983). *Cost effectiveness: A primer*. Beverly Hills, CA: Sage Publications.
- Levin, H.M., Glass, G.V. & Meister, G.R. (1984). *Cost-effectiveness of four educational interventions*. Stanford, CA: Institute for Research on Educational Finance and Governance, Stanford University.
- Monk, D.H. (in press). *Educational finance: An economic approach*. New York: Random House.
- Monk, D.H. (1989). The education production function: Its evolving role in policy analysis. *Educational Evaluation and Policy Analysis*, Spring.
- Murnane, R.J. & Cohen, D.K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Education Review*, 56(1), 17.

- Murnane, R.J., Singer, J.D., & Willett, J.B. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research. *Educational researcher*, 17(6) (August-September), 22-30.
- Rosenholtz, S. (1987). Effective Schools: Interpreting the Evidence. *American Journal of Education*, 93(3), 381-382.
- Rumberger, R.W. (1987). The impact of salary differentials on teacher shortages and turnover: The case of mathematics and science teachers. *Economics of Education Review*, 6(4), 389-400.
- Shepard, L.A. & Kreitzer, A.E. (1987a). *The Texas teacher test*. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April.
- Shepard, L.A. and Kreitzer, A.E. (1987b). *The Texas teacher test*. *Educational Researcher*, 16(6) (August-September), 22-31.
- Shulman, L.S. and Sykes, G. (1986). *A national board for teaching? In search of a bold standard*. Paper prepared for the Task Force on Teaching as a Profession. Carnegie Forum on Education and the Economy.
- Solmon, L.C. & Fagnano, C.L. (1988, April). *Speculations on the benefits of large scale teacher assessment programs or how 78 million dollars can be considered a mere pittance*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- U.S. Department of Education. (1987) *What's happening in teacher testing*. Washington, DC

24. Embracing Contraries: Implementing and Sustaining Teacher Evaluation

Milbrey Wallin McLaughlin

Teachers are evaluated by one means or another in virtually every school district. And in most of those districts, teachers and administrators agree that the activity is ritualistic and largely a waste of time (Stiggins & Duke, 1988; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984; Harris, 1986; Kauchak, Peterson, & Driscoll, 1985). Teacher evaluation generally results in neither more accountability for, nor improvement in, educational practices.

Practitioners and policy makers trying to do better have discovered that the problem of teacher evaluation is not simply or even a question of mandating better evaluation practice. The problem of teacher evaluation has multiple and complex dimensions. One of them is capacity. Are the skills and strategies essential to strong teacher evaluation available to the district? This is a *technical* problem that raises issues of training and program design.

Another problem is that of the relationship between teacher evaluation and other district or school level activities. Is there integration between teacher evaluation and such areas as staff development or curriculum that could benefit from the information generated by a teacher evaluation program? This is a *management* problem that encompasses issues of the institutional (and so individual) significance of teacher evaluation activities.¹

Most discussions of how to carry out more effective teacher evaluation center on these two concerns and debate the merit of particular evaluation plans or the advisability of joining accountability and improvement in the same teacher evaluation strategy. In most deliberations, "The myth lives on that planning for teacher evaluation is simply creating a new instrument" (Harris, 1986, p. 248). As important and difficult as these technical and management problems are, they

AUTHOR'S NOTE: The phrase *embracing contraries* originates with Peter Elbow (1986) as a description of the complexities of teaching and learning. With credit to Elbow, I adopt the term here to signal the issues central to teacher evaluation.

This essay draws heavily on the research and analysis reported in McLaughlin and Pfeifer (1988). Nina Pascia, Edwin Bridges, Robert Calfee, Ann Lieberman, Gary Natriello, and Donald Thomas all provided thoughtful reviews of this chapter in draft; their comments and suggestions prompted clarification and elaboration at several points. I appreciate their feedback.

must wait on the solution to a third problem, one that is fundamental to implementing and sustaining a meaningful teacher evaluation program — establishing a culture for teacher evaluation. This is a difficult problem of *organizational change*.

Establishing a Culture for Teacher Evaluation

If there is an irreducible truth of teacher evaluation, it is: Teacher evaluation will be no more effective than the extent to which teachers support it. An effective teacher-evaluation system assumes candor on the part of teachers and requires opening classroom doors for meaningful observation by evaluators, rather than the dog-and-pony show teachers say most evaluators witness.² An effective evaluation system demands teachers' willingness and ability to act on the outcomes of an evaluation. An effective evaluation system insists on trust between teachers and administrators.

These enabling conditions exist in few school districts. Instead of trust, relationships between administrators and teachers more often reflect them/us divisiveness (Blumberg, 1980; Bridges 1986). Instead of candor, teachers tend to hide problems, and then to hide the fact that they are hiding them. Instead of risk-taking and efforts to work on new classroom strategies, many teachers hold on to established practices, even those acknowledged to be less effective than they might be. In most school districts, the norms and expectations that surround teacher evaluation preclude a meaningful activity. As one long-time observer of teacher-evaluation practices put it: "Because teacher evaluation is complex, threatening, and not well-understood, much of current practice involves 'games' rather than systematic evaluation" (Harris, 1986, p. 12).

The fundamental problem confronting practitioners and policy makers hoping to establish a strong teacher-evaluation program is to change those norms and expectations and to establish a culture that supports evaluation. Teacher evaluation can have power for accountability or for improvement only when the values and behavior it requires are part of the district's institutional fabric and seen as legitimate by teachers. The particular teacher-evaluation system selected by district educators will and should vary from setting to setting (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984; Stiggins & Duke, 1988; McLaughlin & Pfeifer, 1988). One district's programmatic approach to teacher evaluation may be incompatible with the organizational traditions, needs and capacity of another. Nonetheless, there are common elements in the implementation process associated with success in building and sustaining a district culture for teacher evaluation. This chapter focuses on these elements:

- evaluation embedded in a broader improvement effort
- stakeholder involvement
- active commitment of the superintendent

- joint training for administrators and teachers
- resources to support individual needs.

*Evaluation Embedded in
a Broader Improvement Effort*

"The check is in the mail." Teachers accord claims about the value of teacher evaluation just about as much credibility as credit managers grant to assurances about posted payments. Teachers' suspicion and cynicism about the purposes of evaluation have healthy tradition. Inspection and control have characterized teacher-evaluation activities since colonial times (Peterson, 1982). And, at least since the turn of the century, teachers have protested that performance appraisal undermines their professional status and misrepresents the teaching task. In 1916, for example, the president of the National Education Association's Department of Classroom Teachers launched a forceful attack on the so-called efficiency system as a pernicious and wrong-headed effort to standardize the relationship between teachers and students (Peterson, 1982, p. 38). More recent responses sound the same theme: "Negative perceptions about evaluations and programs have often developed because of poor communications and a feeling that some programs are designed to penalize rather than reward teachers" (Southern Regional Education Board, 1986, p. 2).

Districts that have made significant strides in changing the norms and behavior associated with teacher evaluation offer an important lesson. Teachers considered evaluation meaningful and in their professional self-interest only when it is embedded in a broader district effort to improve educational practices.³ Teacher evaluation activities acquire legitimacy and teachers' support when it is viewed not as an end in itself but as instrumental to a highly visible, comprehensive improvement focused effort.⁴

For example, the superintendent in Moraga, California invested heavily in staff-development programs and clinical training for all staff, thereby establishing clear district priority for improvement. Teacher evaluation was addressed only after these staff development activities had been in operation for two years; a strengthened teacher evaluation system became the logical next step. Charlotte-Mecklenberg's innovative Career Development Program, adopted and designed with active teacher involvement, incorporates a strong teacher evaluation component. Toledo's teacher-run effort to improve the quality of the district's teaching force through a strong mentoring and advising program hinges on a substantial teacher evaluation component. In Santa Clara, California and in Salt Lake City, teacher evaluation is a central part of the "shared governance" model initiated to improve the quality of education in the district. Likewise, in another California district, Mountain View-Los Altos, a strong teacher evaluation program was framed in the context of the superintendent's express commitment to professional improvement for all district educators.

In each of these districts, teacher evaluation is part of a broader effort to improve practices and relationships; it derives its legitimacy from this association. Teacher evaluation will be perceived as "good for teachers" only when it has this contextual credibility. When it precedes rather than follows district-wide improvement initiatives (or when it is undertaken in the absence of any significant district-level commitment to improve practice) teacher evaluation most likely will be regarded as threatening and contrary to teachers' professionalism.

Stakeholder Involvement

Stakeholder involvement in the development of new practices and the participants' resulting sense of ownership recurs as a theme in analyses of successful planned change efforts (Fullan, 1982, 1987). Successful teacher evaluation initiatives, which require the most difficult kind of change, change in norms and beliefs, are no different.

Teachers' involvement is an irreducible requirement. The exclusion of teachers from the process perpetuates a them/us schism between administrators and teachers, which is fatal to teacher evaluation and reinforces a view of teacher evaluation as indifferent to teachers' professional expertise and classroom realities. The openness and support required by a meaningful evaluation effort simply will not develop unless teachers have an active role in making decisions about how their performance will be assessed and about the consequences of evaluation.

Districts have involved teachers in a variety of ways and at all points in the process. Santa Clara used a teacher team to review practices around the country and recommend a policy for the district; teachers are central to the district's remediation efforts. Charlotte-Mecklenberg engaged teachers at every step. Teachers in Charlotte single out their involvement as a main reason "why things are so good:"

... the answer [to why new practices are working in Charlotte] is teacher input. We were involved all the way. It is true that mistakes have been made ... and there is a need to be flexible, but the bottom line is that when teachers are being heard, success is possible. (McLaughlin & Pfeifer, 1988, p. 26).

Two lessons about teachers' involvement stand out in the experiences of districts that have been relatively successful in securing teachers' support for evaluation. One is that simply appointing teachers to steering committees or to planning groups is insufficient to generate support. Not all teachers are willing or able to assume this role. Teachers who are not involved in the process may be no more willing to back an evaluation system than are teachers in districts where teachers' participation is not invited.

Likewise, while engagement of the teachers' organization is critical to the success of teacher evaluation in both the short and the long term, it is insufficient

to establish a culture for evaluation.⁵ All teachers in the district, not just organization officials, must be given an opportunity to express their opinion and offer suggestions. Districts have accomplished this through such homely devices as school building suggestion boxes, or such time consuming activities as building level meetings between teachers and district planners and administrators. As the Charlotte-Mecklenberg experience indicates, not all teachers will take advantage of opportunities to comment at various points in the process. But the opportunity to do so must be available and must be seen as meaningful.

The involvement of teachers, in summary, contributes at least three essential factors to the successful development and implementation of a teacher evaluation effort: greater teacher commitment to evaluation and motivation to participate fully; greater tolerance for the inevitable mistakes and flaws in the developing system; advice and assistance for teachers during the course of development that may be crucial to the system's long-term viability.

Still another lesson emerges from the experience of districts that have achieved some success in developing a culture for evaluation: Teachers and administrators are not the only stakeholders in evaluation. Parents, district trustees, and other interested citizen groups also have an immediate interest in the quality of education in the district and so in teacher evaluation. Broader district and community support for an evaluation system can be generated through sessions to orient these groups to the purposes, strategies and expectations of the districts evaluation plan (Harris, 1986, p. 237). Involvement of these stakeholders can create a zone of tolerance to buffer administrators and teachers when things don't go as planned or when the inevitable mistakes are made in the uncertain process of putting a new strategy in place.⁶

Active Commitment of the Superintendent

Enabling teacher evaluation — establishing a culture for evaluation — is not a task that can be delegated to middle management, to a consultant, or to a committee of administrators and teachers. Enlisting the support of teachers for a strong teacher evaluation program requires the active, express and visible commitment of the district superintendent. The superintendent is the spokesperson for evaluation purposes, for the ways in which evaluation will be used and for the priority accorded to it. The superintendent has a particular and nontransferable role to play in establishing the values and expectations that surround teacher evaluation. One important way in which that is expressed is by locating teacher evaluation within a broader improvement effort, thereby signaling that teacher evaluation is undertaken in the service of the broader goal of improvement and professionalism.

Active superintendent involvement is important for other reasons. Development of trust between administrators and teachers requires open channels of communication up and down the district; the superintendent is the ultimate keeper of those open channels by his or her own behavior and by the messages

sent to administrators. Without active channels to convey messages from the district offices to the school and from site personnel to central office staff, the vertical integration essential to coherent district policies and programs is difficult to achieve or maintain.

The superintendent also is essential to promoting the meaningful participation of administrators in various aspects of teacher evaluation — planning, training, and conduct of evaluation. Administrators traditionally are as allergic as teachers to a strong teacher evaluation system — it is time-consuming; it threatens delicate relationships; it competes with other priorities. The superintendent, as final arbiter of professional self-interest in a district, must make it plain through his or her own active commitment that strong teacher evaluation is a priority for the district, that teachers will have a voice in it, and that teacher evaluation serves goals consistent with professional norms and values.

Joint Training for Administrators and Teachers

Research on teacher evaluation points consistently to the need to train administrators in evaluation procedures and skills (Harris, 1986; Stiggins & Duke, 1988; Wise, et al., 1984; McLaughlin & Pfeifer, 1988; Kauchak, Peterson, & Driscoll, 1985; Medley & Coker, 1987).⁷ Administrators require training in observation, in recording classroom activities, in conducting pre- and post-evaluation conferences with teachers and in strategies for following-up on evaluation.⁸ Administrators also require exposure and experiences that supports change in the typical professional culture of school administrators — a perspective which frames evaluation of teachers as a routine, checklist activity undertaken to fulfill bureaucratic requirement rather than to inform individual or institutional improvement efforts.⁹

Especially because of the need to foster new beliefs and understandings about teacher evaluation among administrators as well as equip them with new skills and competencies, training sessions most effective in supporting and enabling to a healthy evaluation system and to a district-wide culture for evaluation involve teachers as well as administrators. Joint training sessions comprise important opportunities to develop common understanding about evaluation, common expectations for the process and outcomes, and common language to talk about all aspects of the process. The inclusion of teachers in evaluation-related training session contributes critically to the development of shared goals, knowledge, and expectations — features central to an evaluation culture. In addition, teachers have important expertise to add. For example, many administrators report that teachers' input was essential to their understanding of classroom realities and to identifying approaches that enhanced the function of evaluation in the district. Teachers are the experts about factors shaping many classroom decisions, about the issues that concern them the most, and about ways to support meaningful response to evaluation findings. This expertise is essential to a healthy program.

Joint training also responds to teachers' needs for training in the details, expectations and procedures of a new evaluation system; it enables teachers to become "thoughtful consumers" and knowledgeable participants in the process (Kauchak, et al., 1985).¹⁰ Joint training sessions serve an important symbolic purpose. They signal that teacher evaluation is a shared responsibility in which all district educators have an important role to play, something undertaken *with* teachers, not done *to* teachers, as part of a comprehensive district initiative.

Resources to Support Individual Needs

Taking evaluation seriously and giving it credibility from teachers' perspective means that the strategy must be *investment centered* not pay-off focused.¹¹ A pay-off oriented evaluation system rewards teachers for past performance and so encourages "safe" observations and problem-hiding; an investment-centered system rewards teachers for undertaking professional development activities and encourages risk-taking and acquisition of new skills. An investment-centered system takes *individual* needs seriously and represents a developmental model of professional growth.¹² In this view, the issues raised by teachers' evaluation and needs for resources probably will differ by classroom situation and by stage of teaching career. Individualized responses to teachers' evaluations allows substantive responses to "good" teachers as well as to teachers needing to make fundamental changes in their classroom practices. Individualized responses to evaluation provides important testimony to improvement and provides a powerful incentive for teachers to approach evaluation with candor and openness. Veteran observers of teacher evaluation schemes insist on this feature. For example, Harris states: "Resources provided for the total evaluation system operation should be *clearly* supplemented with resources for supporting the improvement efforts of individual teachers" (1986, p. 254). Glickman also sees individualized response as essential (1985, p. 273):

To leave classroom teachers alone and unobserved in their classroom, without professional consultation and without school resources tailored to their unique needs, is a statement (intended or not) that teaching is unimportant. The message to teachers is that what is important is keeping your class quiet, your doors shut, and your problems to yourself.

An evaluation system that responds to teachers' individual needs requires the availability of multiple resources for teachers evaluators (McLaughlin & Pfeifer, 1988). Routine district workshops or regular in-service sessions won't fill the bill. Districts committed to investing in teachers' growth have provided resources as diverse as funds for a university course, release time to visit colleagues' classrooms, money for an innovative project, or support for consultants (Wise, et al., 1984; McLaughlin & Pfeifer, 1988). Evaluators, in this instance, become managers of opportunities for teachers' professional development.¹³

The availability of resources to respond to individual's needs serves accountability of the most fundamental kind, accountability rooted in professional norms and values. To this point, districts that take an improvement-focused approach to evaluation demonstrate two important outcomes: one, most teachers are able to make substantial and important changes in their practice given feedback and the opportunity to improve; and two, the majority of those teachers who were unsuccessful in meeting district standards voluntarily left teaching, believing they are been accorded sufficient and fair opportunity to become more effective (Wise, et al., 1984; McLaughlin & Pfeifer, 1988).

Sustaining Teacher Evaluation

Establishing a culture for teacher evaluation is a necessary, difficult and time consuming requirement for implementing a meaningful system. Analysts in the United States and in Canada who have examined teacher-evaluation practices in a variety of settings say that five years is a not unusual amount of time to effect the conditions that enable teacher evaluation (Fullan, 1988; Harris, 1986).

Establishing a culture for teacher evaluation and implementing a promising teacher-evaluation scheme, however, is the beginning not the conclusion of the process. A district culture that supports evaluation and the viability of the strategies selected needs to be sustained if teacher evaluation is not to revert to ritualistic and relatively meaningless activity. The following features are important to sustaining a culture for evaluation. They comprise activities to ensure that evaluation continues to be taken seriously by all actors, to be viewed as fair by teachers, and to make meaningful contributions to educational practices in the district.

Continued Active Involvement of the Superintendent

The superintendent's central role in implementing a strong teacher evaluation plan does not end once a plan is developed and installed. His or her continued and active interest appears to be necessary to keep administrators' attention focused on teacher evaluation as a top priority. In the absence of pressure and support from the top, responsibilities for teacher evaluation are likely to be bumped by competing demands in the hectic life of a school administrator. The experience of Santa Clara Unified School District shows how a diversion of the superintendent's express attention to teacher evaluation can undermine the process — even when it is apparently well-established (McLaughlin & Pfeifer, 1988, p. 22). The superintendent's successful handling of problems associated with fiscal retrenchment and enrollment decline received well-deserved national attention, but the concomitant lapse in the superintendent's express attention to teacher evaluation was reflected down through the system as principals, too,

assigned teacher evaluation a lower priority. The continued, active involvement of the superintendent is essential to maintaining teacher evaluation as a priority for administrators and teachers.¹⁴

Accountability for Evaluation

An accountability strategy to reinforce the superintendent's commitment to evaluation also is necessary to sustaining teachers' support for and administrative attention to teacher evaluation. Two kinds of accountability are important to both the near term and the long term support of teachers. One is some form of checks and balances that promotes accountability of the system itself through an emphasis on fairness, professionalism and expert judgment. For example, the Charlotte evaluation system had provisions for multiple observations, an appeal process in the event of disagreement, and an insistence on extensive documentation of assistance efforts.

Another form of accountability focuses on administrators and holds them responsible for professional and comprehensive teacher evaluation consistent with district guidelines and expectations. For example, in Moraga and Mountain View-Los Altos central office administrators (including the superintendent) read and critique the evaluations and observation reports (McLaughlin & Pfeifer, 1988, pp. 44-46). Administrators' performance as an evaluator also is assessed as part of an administrator's performance review. Holding evaluators accountable for the quality of their evaluation focuses attention on evaluation and reinforces assurances to teachers that evaluation will be taken seriously and carried out in a professional manner.

Strategies for Review and Revision

Teacher evaluation is not an event but a dynamic, evolving process. Just as the need for the superintendent's active involvement continues, so does the need for the review and design sessions of the type that occurred in the planning stages. Information about teachers' perceptions of the process, about the consequences of evaluation, and about implementation issues is required to monitor the process with the purpose of making adjustments in teacher-evaluation policy or practice. The on-going need for revision and adjustment reflects the necessarily incremental process of implementation. For example, Harris (1986, p. 238) advises:

Most new, complex systems will have greater chances for success if some gradual or incremental plan for implementation is utilized. Wholesale adoptions . . . are rarely effective. Any complex operation calls for many changes and time must be allowed for making them with minimum cost, pain, and distress . . . (Study groups) are especially useful when a new system is controversial and when group commitment is essential to success.

Strategies for continuing review and revision also recognize change as a fact of school district life. New classroom realities, new demands on teachers, or changed teacher demographics can generate new issues for teacher evaluation; new actors need training and socialization into the district's evaluation culture.

On-going review and revision also recognizes that the evaluation process itself is developmental. In most of the districts mentioned here, for example, teachers and administrators found that in time their expertise and expectations outdistanced those assumed by initial plans and procedures. Revisions in the process and procedures were necessary to accommodate new levels of a competence, capacity and expectations.

Integration of Individual and Institutional Activities

Teachers' support for evaluation and sustaining those beliefs in the long-term requires more than trust between teachers and administrators. It also turns on participants' perception of the process as useful and contributing to a more effective educational program. Individual learning and professional development is just one component of this assessment. The reflective individual practice engendered by a strong evaluation system can underwrite reflective practice at the institutional level. In this view, the outcomes of individual evaluations have implications not only for the teacher involved, but also for the organization. What does an individual's evaluation say, for example, about district staff development activities? About procedures for curriculum development and review? About personnel policies? About principals' relations with the central office?

Integration of individual and institutional activities serves to promote and sustain culture for evaluation in a number of ways. It enables the system as well as the individual to learn as a consequence of teacher evaluation. For example, Mountain View-Los Altos' staff development offerings were revised substantially as a consequence of teachers' comments during the evaluation process. Charlotte-Mecklenberg rethought its strategies for hiring new teachers based on evaluation evidence. Lake Washington changed its promotion and review policies for administrators. District-level responses such as these provide evidence that the district takes evaluation seriously. They also enhance teachers' sense of professional safety; instead of seeing a teacher's performance only in individual terms, district officials interpret it against district choices and activities. Teachers' candor thus is rewarded not only with resources to promote their professional development, but also by changes in district practices and policies consistent with system improvement goals. Consequently, teachers can see how their participation in and support for evaluation can make the district a better place for students and teachers. Integration of individual and institutional responses thus is another powerful strategy for reinforcing an ethos of "we" perspective among teachers and administrators and so for sustaining an evaluation culture. It also underscores and sustains the two ways in which an improve-

ment oriented evaluation system contributes to the effectiveness of the school system — by fostering the improvement of teaching and by creating the conditions that nourish its own improvement.¹⁵

Embracing Contraries

Contraries take center stage in the teacher evaluation story. Such contraries as

- accountability/improvement
- bureaucratic control/professional autonomy
- superior/subordinate
- problem hiding/problem solving
- individual needs/institutional goals

bedevil proponents of strong teacher evaluation and defeat efforts to implement meaningful teacher programs in most districts. Yet they also are essential to a productive, vital school district. Both accountability and improvement are basic goals; both teachers and administrators have indispensable functions to fulfill; bureaucratic rules must coexist with professional norms because education is a public function subject to review and oversight by various agencies of government. However, failure to confront the contraries intrinsic to teacher evaluation ensures the continued failure of teacher evaluation initiatives. A constructive response requires strategies to blend and blur (rather than exacerbate) these contraries. An evaluation system located within a broader improvement effort can foster accountability through its investment in professional norms and values; these contraries thus become mutually reinforcing rather than competing goals. Similarly, the contraries that fuel confrontations between bureaucratic forms of control and norms of teacher professionalism can be softened when evaluation resources underwrite individual efforts to improve.

The cold war between administrators and teachers can be dampened if not extinguished by common training and collaboration for teacher evaluation. Strategies for review, revision and integration of individual and institutional activities can provide incentives for all actors — central office staff, principals and teacher — to take evaluation seriously, to be candid about problems, to communicate, and to invest in efforts to get better.

Implementing and sustaining an effective teacher evaluation system means embracing contraries, understanding the roots of their friction, and seeking ways to harmonize them in a culture for evaluation. Once this organizational problem has been tackled, solutions to the technical and management problems of evaluation will be easier to realize and maintain.

Notes

1. Donald Thomas, former superintendent of schools for Salt Lake City comments at this point that the problem of teacher evaluation is also a problem of communication. Thomas writes: "A third problem is the misunderstanding that exists about the purpose of evaluation. Is it a process for personnel decisions? Is it a program to assist poor teachers? Is it a way to identify superior teachers? Is it something that satisfies all these purposes? This is a communication problem and few school districts have a clear-cut, focused purpose for evaluation and a process which is more than classroom observation" (personal communication).

2. Students often are enlisted in the theater of teacher evaluation. David Tyack tells the story of the rural school, for example, where the imminent arrival of an evaluator to a teacher's classroom was heralded by a student from the classroom currently under observation. "May we borrow the colored chalk?" This question signalled that the evaluator was on the way; the school had no colored chalk.

3. In his review of this chapter, Gary Natriello offers a caveat to this statement: "I think that you are overstating the resistance of teachers to teacher evaluation. Many teachers are delighted to participate in evaluation because it is one of the few times when someone pays attention to them. The isolation of teaching is so great that evaluation, even when it is negative, is a welcome contact with other adults. This should be encouraging to those interested in implementing teacher evaluation programs. Teacher will show interest in such systems even when they are flawed. However, teachers are likely to work to improve such systems even when they don't meet the rather stringent standard specified in your statement."

Our experience suggests that the interests and incentives Natriello highlights to characterize many teachers' response to teacher evaluation, but that nonetheless the district's history with evaluation and teachers' perceptions of its merit and purposes are powerful influences on teachers' willingness to become involved in a meaningful way.

4. This experience and the examples that follow is reported in McLaughlin and Pfeifer (1988) and in Wise, et al. (1984).

5. For examples of active teacher organization involvement, see the Lake Washington and Toledo cases in Wise, et al., 1984.

6. See the Charlotte-Mecklenberg case as reported in McLaughlin and Pfeifer (1988) for an example of a broad-based district and community support in the face of some implementation disappointments.

7. These studies also show that teachers' perceptions of administrators lack of competence to evaluate their practices reflected grade level concerns. Elementary teachers complain that their administrator either has little instructional competence, or no experience at a particular grade level. Secondary school teachers challenge the ability of their administrator evaluator to make judgments about their competence in specific subject areas.

Across all grade levels, Kauchak et al. (1985, p. 34) found that a major factor in teachers' positive perceptions of supervision classroom visits was the expertise of the principal.

8. Here Donald Thomas cautions: "Be careful not to limit evaluation to classroom observation. Evaluation includes much more — analysis of preparation, student files and subject matter knowledge. It also includes analysis of student performance, achievement of objectives, and attendance patterns as well as whatever else is of interest or beneficial to the parties involved. Observation is only a small part of a total comprehensive evaluation program" (personal communication).

9. Thanks to Gary Natriello for underscoring the need to change the culture of school administrators as well as that of teachers in the area of evaluation. Natriello points out that administrators are ill-equipped to undertake evaluation as envisioned here not only because of the routines established in most school districts, but also because of the socialization they receive in administrator training programs, where according to Natriello they "are too often trained to be managers, not educators" (personal communication).

10. Teachers serving as peer evaluators have the same training requirements as administrators. Simply being a teacher does not provide the necessary grade or subject matter expertise, or the interpersonal and procedural skills required by a sound evaluation system.

11. See Rosabeth Moss Kanter (1983) for development of the concept of an "investment centered" approach to management and employee appraisal. See also McLaughlin and Yee (1988) for elaboration of this notion in the context of teachers' careers.

12. Glickman (1985) and Harris (1986) present evaluation and supervision models which elaborate a developmental approach to performance appraisal and teacher improvement efforts.

13. Harris cautions that individual plans must be limited to those improvement activities and objectives that the district is able to willing to support (1985, p. 254). Unreasonable or insupportable expectations quickly erode teachers' belief in the evaluation system.

14. Superintendents have used a variety of strategies for maintaining attention to evaluation — making sure it was on the agenda of the annual administrators' evaluations; participating actively in the review and revision of the district's evaluation plan.

15. Thanks to Gary Natriello for clarifying this point.

References

- Blumberg, A. (1980). *Supervisors and teachers: A private cold war*. (2nd ed.) Berkeley, CA: McCutchan.
- Bridges, E.M. (1986). *The incompetent teacher*. Philadelphia: The Falmer Press.
- Elbow, P. (1986). *Embracing contraries: Explorations in learning and teaching*. New York: Oxford University Press.
- Fullan, M.G. (1982). *The meaning of educational change*. New York: Teachers College Press.
- Fullan, M.G. (1987). *Implementing educational change: What we know*. Ontario Institute for Studies in Education. Prepared for the World Bank seminars, Planning for the Implementation of Educational Change.
- Fullan, M.G. (1988). Performance appraisal and curriculum implementation research, (pp. 71-83). In E.S. Hickcox, S.R. Lawton, K.A. Leithwood, & D.F. Musella (Eds.). *Making a difference through performance appraisal*. Toronto: Ontario Institute for Studies in Education Press.
- Glickman, C.D. (1985). *Supervision of instruction: A developmental approach*. Boston: Allyn & Bacon.
- Harris, B.M. (1986). *Developmental teacher evaluation*. Boston: Allyn & Bacon.
- Kanter, R.M. (1983). *The change masters*. New York: Simon & Schuster.
- Kauchak, D., K. Peterson, & A. Driscoll (1985). An interview study of teachers' attitudes toward teacher evaluation practices. *Journal of Research and Development in Education*, 19, (1), 32-37.
- McLaughlin, M.W. & R.S. Pfeifer (1988). *Teacher Evaluation: Improvement, accountability and effective learning*. New York: Teachers College Press.
- McLaughlin, M.W. & Yee, S.M. (1988). School as a place to have a career (pp. 22-44). In A. Lieberman, (Ed.). *Building a professional culture in schools*. New York: Teachers College Press.
- Medley, D.M. & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, 80(4), (March/April), 242-247.
- Peterson, C.H. (1982). *A century's growth in teacher evaluation in the United States*. New York: Vantage Press.
- Stiggins, R.J. & Duke, D. (1988). *The case for commitment to teacher growth: Research on teacher evaluation*. Albany: State University of New York Press.
- Southern Regional Educational Board (1986). *Career ladder clearinghouse*. Atlanta: Southern Regional Educational Board.
- Wise, A.E., Darling-Hammond, L., McLaughlin, M.W. & Bernstein, H.T. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica: The Rand Corporation.

25. Using the Personnel Evaluation Standards to Improve Teacher Evaluation

Daniel L. Stufflebeam

James R. Sanders

This chapter is an overview of the new *Personnel Evaluation Standards*, with special attention to how they can help to improve evaluations of teachers. The *Standards* were developed by the Joint Committee on Standards for Educational Evaluation,¹ whose 16 members were appointed by 14 of the most influential professional associations² concerned with education. Published in 1988 by Sage Publications, *The Personnel Evaluation Standards* are designed to assist school districts, universities, and other educational agencies to examine and improve their systems for evaluating the qualifications and performance of professional personnel.

Background of the Standards

The Joint Committee on Standards for Educational Evaluation was established in 1975 for the purpose of developing standards for educational program evaluations. Its original project culminated in the McGraw-Hill publication, *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee, 1981).

At its first meeting in 1975, the Joint Committee explicitly agreed *not* to develop standards for evaluations of teachers or other personnel. The Committee represented (and still does) diverse and sometimes conflicting constituencies, for example, teachers' unions and associations of school board members, administrators, and testing specialists. In 1975, these groups were cautiously willing to collaborate in determining how program evaluations should be judged, but they lacked the mutual trust and cohesiveness required to collaborate in writing standards for the volatile domain of personnel evaluation. It seems fair to say that

AUTHORS' NOTE: A critique of a prior draft of this article by Kenneth Strike was helpful in finalizing this chapter.

the teacher representatives on the Committee feared that the psychologists and researchers would promote more testing of teachers and that the psychologists and researchers worried that the teachers' union representatives would pursue a political rather than professional agenda if the Committee tried to set standards for personnel evaluation systems. Also, in 1975, program evaluation was high on the list of educational problem areas in the United States needing immediate attention, but there was little concern from any sector for improving educational personnel evaluation. Thus, the Committee started its collaborative effort by limiting its objectives to the domain of program evaluation.

However, in 1984, the Joint Committee changed its collective mind and decided explicitly and unanimously to develop *The Personnel Evaluation Standards*. There were three important reasons for the reversal.

First, the Committee had succeeded in developing the program evaluation standards and had developed an effective process for confronting controversial issues and reaching consensus on standards. In the meetings of the Committee, there had been much conflict and many heated exchanges, but, in addressing issues, the Committee stayed its course, followed due process, obtained external opinions, and reached democratically determined decisions. Ultimately, it found much agreement and learned to value its diversity. Also, the Committee had become incorporated so it could continue to promote the use of the program-evaluation standards, periodically review and revise them, and, if needed, develop other sets of standards for other domains of educational evaluation. In 1984, it seemed clear that the Committee had matured sufficiently so that it could deliberate professionally, collaboratively, and effectively on issues pertaining to standards for personnel evaluation.

Second, the Committee concluded that standards were needed to diagnose problems in present personnel evaluation systems. By 1984, widespread consensus had developed in the United States that personnel evaluation practices in education were in a sorry state and needed to be reformed. Several national reports underscored the need to increase and improve evaluation of teacher candidates, teachers, and other education personnel (Boyer, 1983; Goodlad, 1983; NCEE: *A Nation At Risk*, 1983; NSF: *Educating Americans for the 21st Century*, 1983; Peterson: *Making the Grade*, 1983; NEA, 1983). Soar, Medley, and Coker (1983, p. 246) concluded that teachers' resistance to evaluation was reasonable, since evaluations then in use were "subjective, unreliable, open to bias, closed to public scrutiny, and based on irrelevancies"; and Darling-Hammond, Wise, and Pease (1983) echoed this position, saying that most extant systems were illogical, simplistic, unfair, counter productive, or simply unproductive. In light of these charges, school districts and state education departments throughout the United States needed to examine and improve their personnel evaluation systems.

Third, the Committee thought that standards were also needed as a basis for evaluating personnel evaluation reform proposals. The 1980s had seen a number of projects aimed at reforming teacher evaluation, using such approaches as

merit pay, payment by results, career ladders, master teacher plans, peer review, clinical observation, and objective tests. While some of the projects were carefully conceived and rigorously developed, many others were hastily designed, hugely expensive, highly controversial, legally vulnerable, and/or of dubious merit.

For the above reasons, the Joint Committee in 1984 decided, with the concurrence of its sponsoring organizations, that it was qualified to undertake the Personnel Evaluation Standards project and that it should do so with all due haste. It spent the next few months doing extensive background analysis and planning.

Development of the Standards

In 1985, with an initial grant from Lilly Endowment, Inc., the members of the Joint Committee agreed on a structure and plan for developing the standards.³ The structure included an outline of the projected standards book, a list of issues to be converted into a first set of standards, a format for writing drafts of each standard, and plans for involving a wide range of constituencies in the process.

During the first year of the project, many individuals were involved through writing and review panels, symposia, a blue-ribbon validation panel, and other vital activities. In 1985, a first draft of the Standards was produced and critiqued by international and national review panels and the validation panel. In 1986, the critiques were used to develop a second draft, and, in 1987, the second draft was field tested and subjected to national hearings. The testimony received from 24 individuals at the hearings, 36 field test reports, and a report from the validation panel were used to develop the final publication.

Personnel Evaluations for Which the Standards were Developed

The Joint Committee defined a comprehensive domain of application for the *Standards*. Table 25.1 depicts the general evaluative context for which the *Standards* were developed. The table portrays personnel evaluation as an integral part of societal and institutional efforts to prepare, engage, and develop educational personnel. The main horizontal dimension includes three systems called *preparation* (e.g., teacher education), *practice* (e.g., school teaching), and *continuing education* (e.g., study leaves and inservice education). The vertical dimension divides each system into *entry* activities (e.g., selection of students for a principalship program), *participation* (e.g., tenure reviews), and *exit* (e.g., reduction in force decisions). The second horizontal dimension denotes *evaluations and decisions* that are involved in the entry, participation, and exit stages of each personnel system.

The entries in the cells of the matrix reveal that a wide range of evaluations and decisions are involved in educational personnel work. Moreover, the evaluations are of three different types. Some of them are program evaluations, e.g., evaluations of recruitment programs. Others, such as evaluations of students' mastery of college courses, are student evaluations. Finally, the great majority of the evaluations identified in the matrix fit the common view of personnel evaluation, i.e., assessments of the qualifications and performance of individual educators for certification, selection, staff development, tenure, and promotion. The full range of evaluations depicted in the matrix are important to effectively staffing educational institutions, and all such evaluations should adhere to appropriate professional stands.

The new *Personnel Evaluation Standards* are focused on the educational personnel evaluations denoted by the matrix. Those evaluations and decisions on which the Personnel Standards are primarily focused are starred (*) and appear in **bold type**. It is interesting and important that given categories of evaluation often provide information for use in making different decisions. This is especially the case for performance reviews that provide information for decisions about tenure, promotion, merit pay, and other awards, as well as counseling for staff development. One key requirement of the *Standards* is that all parties to an evaluation must enter into it with a clear idea of how the information to be collected will be used.

As evident in Table 25.1, the Personnel Evaluation Standards are intended to have broad application. Educational institutions of all types share needs for sound evaluations for entry into professional training, certifying competence, defining roles within the institution, selecting from job applicants, monitoring and providing feedback about performance, counseling for staff development, determining merit awards, and making decisions about tenure, promotion, and termination. The Personnel Evaluation Standards are proposed as a single set of standards that educators and their constituents can use to plan or evaluate evaluation systems for the full range of personnel actions shown in Table 25.1.

The Joint Committee intends, further, that the *Standards* be applied to all professional roles in schools, post secondary institutions, and other institutions that have a primary responsibility to educate. They also intend that the *Standards* will be applicable to a broad range of techniques, including observation, interview, applied performance tests, licensure tests, professional skills tests, development of a portfolio, supervisor assessment, peer assessment, and student assessment.

Guiding Definitions

The Joint Committee has defined personnel evaluation as *the systematic assessment of either or both a person's performance and qualifications in relation to a role and some specified and defensible institutional purpose*. It has

defined a standard as "a principle commonly agreed to by people engaged in the professional practice of evaluation for the measurement of the value or the quality of an evaluation" (Joint Committee, 1981, p. 12).

The Committee presented *The Personnel Evaluation Standards* at the level of elaborated general principles, with a wide range of illustrations constructed to help users see how to apply the standards to the various types of personnel evaluations seen in Table 25.1. The Committee emphasized that general principles are adequate for providing direction for improvement, and that they avoid oversimplifications and leave room for creative and locally responsive evaluation procedures. Even so, the Committee provided practical suggestions on how to meet each standard.

Four Requirements of Sound Evaluations and the Associated Standards

Sustaining the position it adopted in its program evaluation standards, the Joint Committee grounded its development of *Personnel Evaluation Standards* in a fundamental proposition. It is that all evaluations should have four basic attributes: *propriety*, *utility*, *feasibility*, and *accuracy*. Simply stated, the Joint Committee developed substantive standards and practical guidelines that educators can use to examine the extent to which any personnel evaluation system possesses these four attributes.

In this brief chapter the authors can only provide a cryptic summary of the standards; in full form, each standard contains extensive definition, detailed guidelines, lists of common errors to watch for and avoid, and two illustrative cases. Proper use of the *Personnel Evaluation Standards* requires a thorough reading of the complete text of each standard. Given this caveat, stated below are the Committee's definitions of the four attributes of a sound evaluation, with each definition followed by a summary listing of the applicable standards.

Propriety Standards

The Propriety Standards are aimed at ensuring that the rights of persons affected by an evaluation system will be protected, including students, teachers, professors, administrators, and evaluators. These standards reflect the fact that personnel evaluations may violate or fail to address certain ethical and legal principles. The primary principle is that schools exist to serve students; therefore, personnel evaluations should concentrate on determining whether educators are effectively meeting the needs of students. Overall, the Propriety Standards require that evaluations be conducted legally, ethically, and with due regard for the welfare of students, other clients, and educators.

It is interesting that the Joint Committee deliberately placed the Propriety Standards first among the four sets of standards. They did so to emphasize that

Table 25.1 Types of Evaluations and Decisions Involved in Preparing, Deploying, and Developing Professional Educators

Educational Personnel Systems						
Stages of Involvement	Preparation		Practice		Continuing Education	
	Evaluations	Decisions	Evaluations	Decisions	Evaluations	Decisions
Entry	Evaluations of supply and demand	Assigning priorities and allocating funds to specialized training programs	Evaluations of staffing needs	Definitions of jobs. Decisions to fill certain job vacancies	Correlated assessments of institutional and staff needs	Deciding on continuing education offerings/opportunities.
	Evaluations of recruitment programs	Determining how the programs should be changed or strengthened	Evaluative of recruitment programs	Determining how the programs should be changed or strengthened	Assessments of the needs and achievements of individual staff members	*Deciding whether to approve applications for study leaves, sabbatical leaves and for special grants
	*Assessments of applicants	*Selection of staff	*Evaluations of applicants	*Selection of staff members		
Participation	Intake Evaluations	Determining student programs	*Correlated evaluations of jobs and incumbents' qualifications	*Updating of job definitions	Intake Evaluations	Designing individualized continuing education programs
	Evaluations of students' mastery of course requirements	Assigning course grades	*Reviews of job performance and special achievements	Deciding whether to remove or continue probationary status or to terminate	Program reviews	Providing feedback to guide the learning process
	Cumulative progress reviews	Counseling for remediation Revising program plans Referrals to a termination review	Grievance hearings	*Tenure *Promotion *Merit Pay *Counseling for staff development *Honors (awards) *Recertification Rulings on the grievances		
Exit	Final evaluations of students' fulfillment of their programs	Graduation decisions	*Correlated evaluations of finances, staffing needs, seniority of present staff and options for down-sizing	*Reduction in Force decisions	Evaluations of participants' achievements in continuing education experiences	Deciding whether given applicants should be rewarded with future grants and/or leaves
	*Evaluations of qualifications to practice given educational roles	*Certification *Licensing	*Evaluations of performance and/or investigations of charges	*Deciding whether to terminate *Deciding whether to withdraw licenses or certificates	Evaluations of qualifications to practice given educational roles	Certification Licensing New assignments

*Those evaluations and associated decisions which are of most concern in the Standards for Evaluations of Educational Personnel

evaluation systems must, above all, help to ensure that students and other clients of educational services are well served and that evaluation systems must provide for orderly dismissal of staff members who persist in providing poor service. In fact, this point is the essence of the first Propriety Standard, called *service orientation*.

The summary statements for the five Propriety Standards are as follows:

P-1 Service Orientation

Evaluations of educators should promote sound education principles, fulfillment of institutional missions, and effective performance of job responsibilities, so that the educational needs of students, community, and society are met.

P-2 Formal Evaluation Guidelines

Guidelines for personnel evaluations should be recorded and provided to employees in statements of policy, negotiated agreements, and/or personnel evaluation manuals, so that evaluations are consistent, equitable, and in accordance with pertinent laws and ethical codes.

P-3 Conflict of Interest

Conflicts of interest should be identified and dealt with openly and honestly, so that they do not compromise the evaluation process and results.

P-4 Access to Personnel Evaluation Reports

Access to reports of personnel evaluation should be limited to individuals with a legitimate need to review and use the reports, so that appropriate use of the information is assured.

P-5 Interactions with Evaluatees

The evaluation should address evaluatees in a professional, considerate, and courteous manner so that their self-esteem, motivation, professional reputations, performance, and attitude toward personnel evaluation are enhanced or, at least, not needlessly damaged.

Utility Standards

The Utility Standards are intended to guide evaluations so that they will be informative, timely, and influential. Especially, these standards require that evaluations provide information useful to individuals and to groups of educators in improving their performance. The Utility Standards also require that evaluations be focused on predetermined uses, such as informing selection and promotion decisions or providing direction for staff development, and that they be conducted by persons with appropriate expertise and credibility. In general, these standards view personnel evaluation as an integral part of an institution's ongoing effort to recruit outstanding staff members, and, through timely and relevant evaluative feedback, to encourage and guide them to deliver high quality service.

The Utility Standards should be especially welcome to educators who see their institution's performance review system as only ritualistic and not helpful,

or, worse, demoralizing and counterproductive. By applying the Utility Standards, an institution would be guided to clarify intended uses of its evaluation system and do whatever is required to ensure that the system addresses relevant questions, issues useful reports, and provides direction for improvement. The main point of the Utility Standards is to insure that evaluations contribute constructively to helping educators deliver excellent service.

The summary statements for the five Utility Standards are as follows:

U-1 Constructive Orientation

Evaluations should be constructive, so that they help institutions to develop human resources and encourage and assist those evaluated to provide excellent service.

U-2 Defined Uses

The users and the intended uses of a personnel evaluation should be identified, so that the evaluation can address appropriate questions.

U-3 Evaluator Credibility

The evaluation system should be managed and executed by persons with the necessary qualifications, skills, and authority, and evaluators should conduct themselves professionally, so that the evaluation reports are respected and used.

U-4 Functional Reporting

Reports should be clear, timely, accurate, and germane, so that they are of practical value to the evaluatee and other appropriate audiences.

U-5 Follow-Up and Impact

Evaluations should be followed up, so that users and evaluatees are aided to understand the results and take appropriate actions.

Feasibility Standards

The Feasibility Standards emphasize that personnel evaluations are conducted in institutional settings that have limited resources and are influenced by a variety of social, political, and governmental forces. Accordingly, the Feasibility Standards call for evaluation systems that are efficient, easy to use, adequately funded, and politically viable.

The summary establishments for the three Feasibility Standards are as follows:

F-1 Practical Procedures

Personnel evaluation procedures should be planned and conducted, so that they produce needed information while minimizing disruption and cost.

F-2 Political Viability

The personnel evaluation system should be developed and monitored collaboratively, so that all concerned parties are constructively involved in making the system work.

F-3 Fiscal Viability

Adequate time and resources should be provided for personnel evaluation activities, so that evaluation plans can be effectively and efficiently implemented.

Accuracy Standards

Accuracy, the fourth group, includes those standards that determine whether an evaluation has produced dependable information about relevant qualifications or performance of an educator. These standards require that the obtained information be technically defensible and that the conclusions be linked logically to the data. The overall rating of the personnel evaluation against the Accuracy Standards gives a good idea of the evaluation's overall validity.

These standards are especially germane to arguments concerning which variables are appropriate for judging a teacher's qualifications or performance. The position underlying these standards is that the variables must be derived from a valid description of the person's job. Simply showing that a personal characteristic — such as teaching style, quantitative aptitude, or race — is correlated with student achievement is not justification for using the characteristic to measure and judge a teacher. As Scriven (1988) has argued, to do so not only risks prejudicial treatment of individuals, but, since the correlations are based on group data and are never perfect, such practice also produces invalid assessments of persons who rate low on the variable but do well on the job, or vice versa. Our field tests clearly indicated that many personnel evaluation systems need to be improved in how well they define jobs, how effectively they consider environmental influences, how validly they measure job qualifications and performance, and how effectively they control for various kinds of bias.

The summary statements for the eight Accuracy Standards are as follows:

A-1 Defined Role

The role, responsibilities, performance objectives, and needed qualifications of the evaluatee should be clearly defined, so that the evaluator can determine valid assessment criteria.

A-2 Work Environment

The context in which the evaluation works should be identified, described, and recorded, so that environmental influences and constraints on performance can be considered in the evaluation.

A-3 Documentation of Procedures

The evaluation procedures actually followed should be documented, so that the evaluatees and other users can assess the actual, in relation to intended, procedures.

A-4 Valid Measurement

The measurement procedures should be chosen or developed and implemented on the basis of the described role and the intended use, so that the inferences concerning the evaluatee are valid and accurate.

A-5 Reliable Measurement

Measurement procedures should be chosen or developed to assure reliability, so that the information obtained will provide consistent indications of the performance of the evaluatee.

A-6 Systematic Data Control

The information used in the evaluation should be kept secure, and should be carefully processed and maintained, so as to ensure that the data maintained and analyzed are the same as the data collected.

A-7 Bias Control

The evaluation process should provide safeguards against bias, so that the evaluatee's qualifications or performance are assessed fairly.

A-8 Monitoring Evaluation Systems

The personnel evaluation system should be reviewed periodically and systematically, so that appropriate revisions can be made.

Applying the Standards to Teacher Evaluation Systems

The *Personnel Evaluation Standards* provide a carefully developed and widely endorsed basis for assessing and improving teacher evaluation systems. We think it is reasonable to expect school districts, universities, and other educational agencies to use them as a checklist of things to consider in improving their current evaluation systems. Also, the Joint Committee strongly recommends that the Standards be used to examine new and proposed evaluation systems, such as state and national evaluation programs for certifying and licensing educators.

Steps of use in applying the *Standards* were recommended by Stufflebeam and Brethower (1987) and adapted by the Joint Committee for inclusion in the *Personnel Evaluation Standards* book. A summary of the adapted steps, with a few additions, is as follows:

1. Study the Standards

- Consider adopting the *Standards* as the basic reference by which to examine and promote quality in personnel evaluation.
- Make copies of the *Standards* available for study by instructors, administrators, and other interested parties.
- Conduct workshops aimed at teaching the *Standards* and illustrating their use.
- Appoint a Study Group on Applying the Personnel Evaluation Standards. Ideally, it should reflect the composition of the Joint Committee on Standards for Educational Evaluation (i.e., teachers, evaluators, administrators, policy board members, curriculum developers, personnel specialists, psychologists, research and testing personnel, counselors, and specialists in educational law).
- Charge the Study Group to apply the *Standards* to the personnel evaluation system, to report which standards are met and not met, to identify deficiencies to

be overcome, and to issue recommendations for strengthening or replacing the system.

2. Clarify the Purposes of the Evaluation System by determining:
 - Whose work is to be evaluated?
 - Why should the evaluations be done?
 - Who will use the findings?
 - What decisions will be determined or affected and/or what types of actions are evaluatees and managers expected to take in response to evaluation reports?
 - Should the evaluation(s) focus on qualification, performance, and/or effectiveness?
 - What impact is the evaluation system intended to have?
3. Describe the Personnel Evaluation System to Be Examined
 - Assemble relevant documents (e.g., personnel policies, negotiated agreements, job descriptions, letters of appointment, rating and reporting forms).
 - Develop a written description of the evaluation system (delineating how the evaluations are staffed, the qualifications of the evaluators, the pertinent policies, the evaluation purposes and questions, the measurement variables and procedures, the procedures for organizing and analyzing data, the reporting format and schedule, the uses of findings, the management system, and the practices in monitoring and evaluating the evaluation system).
4. Apply the *Standards*
 - Engage the Study Group on applying the *Standards* in assessing the extent to which they believe each standard has been satisfied.
 - For each standard the study group should list strengths and weaknesses of the personnel evaluation system and make an overall decision whether the standard is met, partially met, or not met.
5. Decide What to Do About the Results
 - Engage the Study Group to discuss the results and develop recommendations for improving the system.
 - Share the results with the professional staff and board of the institution and obtain their input.
 - Develop a general plan for the institution to use in improving the evaluation system.

The proceeding is an outline of some general steps to follow in using the *Standards* to examine and redesign a personnel evaluation system. The process is very workable and the *Standards* applicable, as evidenced by the field tests of the draft *Standards*. Those tests showed that institutional committees can work through a process like the one described above within a three to six months period and as a result produce a shared plan for improving a personnel evaluation system. For an excellent example of such an application, see the article entitled, "Review of Personnel Evaluation Systems" (Reineke, Willeke, Walsh, & Sawin, 1988) that appeared in a recent issue of the *Journal of Educational Personnel Evaluation*. In that article, the authors reported how they used the *Standards* to

review and revise the teacher evaluation system employed by the Lincoln, Nebraska Public Schools.

Conclusion

In this chapter we have presented the response of the education profession to a critical need in education, the need for professionally defined and widely endorsed standards that could be used to judge personnel evaluation systems, including especially the systems used to examine the qualifications and performance of teachers. Our experiences thus far in applying the *Standards* have led us to these conclusions concerning the use of the *Standards* to assess and improve systems for evaluating teachers or other personnel:

1. The Service Orientation (Standard P-1) and Constructive Orientation (Standard U-1) of a personnel evaluation system are counter balancing values that must be invoked consistently if evaluations are to assure that students receive competent service and that individual educators are continually guided to improve their skills and services.
2. The Defined Role standard (A-1) is a fundamental concern in designing or evaluating any personnel evaluation system. The validity of criteria and data used in any personnel evaluation must reference a valid definition of the evaluatee's role.
3. Unless the *Standards* are systematically applied, evaluation committees may commit many serious errors, e.g., allowing conflict of interest to influence results, wasting time and resources in gathering data that won't be used, engaging untrained evaluators, allowing political factors to distort process and findings, or producing a controversial or legally vulnerable system.
4. A clear understanding of, and commitment to, the use of the *Standards* by all concerned can foster clear and accepted evaluation guidelines, collaborative development of evaluation procedures, predetermined uses of given evaluations, controlled and appropriate access to personnel evaluation files, clear and timely reporting, use of results to counsel and assist evaluatees, search for and control of bias, and periodic review and improvement of the evaluation system.
5. Collaborative and professional planning and implementation of the personnel evaluation system will help to ensure commitment to its credibility, propriety, and utility. Allow time to engage in reflective discussion with interested stakeholders.
6. Ability to assess performance of teachers and other educators accurately and effectively takes a clear institutional commitment in the form of explicit policies and guidelines, a significant investment of necessary resources and time, effective training of evaluators, consistency and equity in application of the process, and constructive use of findings. Institutions must invest adequately in personnel evaluation systems if those systems are to produce valid and reliable information and have a positive impact on performance.
7. Personnel evaluations often lead to legal proceedings over the issue of basic due process requirements. Many of the requirements of basic due process will be met if the *Standards* are followed. The Joint Committee notes that the *Functional*

Reporting standard should be of assistance in meeting the notice requirement of due process; that adherence to the *Defined Roles* standard enhances the defensibility of a determination that an employee failed to meet her or his responsibilities; and that attending to the Accuracy Standards, especially *Practical Procedures*, *Valid Measurement*, and *Reliable Measurement* helps to meet some of the procedural requirements of due process.

8. Personnel evaluation systems must be subjected to periodic review and development. Policies, contracts, laws, evaluation technology, job assignments, key actors, teacher morale, and other conditions can change and consequently may require changes in the evaluation system. The *Standards* provide a comprehensive checklist of principles, issues, and recommendations to consider when conducting such reviews.

References

- Boyer, Ernest L. (1983). *High school*. New York: Harper and Row.
- Darling-Hammond, L., Wise, A.E., & Pease, S.R. (1983). Teacher evaluation in the organization context: A review of the literature. *Review of Educational Research*, 53, 285-328.
- Goodlad, J.I. (1983). *A place called school*. New York: McGraw-Hill.
- Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York: McGraw-Hill.
- Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators*. Newbury Park, CA: Sage Publications.
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative of educational reform*. Washington, DC: U.S. Government Printing Office. (No. 065-000-00177-2).
- National Education Association. (1983). *The N.E.A. on master teacher/merit pay plans*. Washington, DC: National Education Association.
- National Science Foundation. (1983). *Educating Americans for the 21st century*. Washington, DC: National Education Association.
- Peterson, P.E. (1983). *Making the grade*. New York: The Twentieth Century Fund Task Force on Federal Elementary and Secondary Education Policy.
- Reineke, R.A., Willeke, M.J., Walsh, L.H., & Sawin, C.R. (1988). Review of personnel evaluation systems: A local application of the standards. *Journal of Personnel Evaluation in Education*, 1, 373-378.
- Scriven, M. (1988). The new crisis in teacher evaluation: The improper use of "research-based" indicators. *Professional Personnel Evaluation News*, 1(1), 4-8.
- Soar, R.S., Medley, D.M., & Coker, H. (1983). Teacher evaluation: A critique of currently used models. *Phi Delta Kappan*, 65, 239-246.
- Stufflebeam, D.L., & Brethower, D.M. (1987). Improving personnel evaluations through professional standards. *Journal of Personnel Evaluation in Education*, 1, 125-155.

Index

- Abstract reasoning, student achievement and, 398
- Academic freedom principle, 358
- Accountability, 25
- democratic societies and, 364
 - evaluation and, 116, 159, 411
 - portfolios and, 253
 - professional development and, 127-130
 - professionalism and, 365-366
 - rating scales and, 198
 - Teacher-Evaluation Cycle and, 166-167
- Accounting, unlicensed practitioners in, 64
- Achievements, teacher
- evidence of merit and, 79
 - interviews and, 94-95
 - need for, 123
- Action, classroom observation and, 178
- Activity checklists, 197, 221
- Adaptation skills, 140
- Administrators
- anti-intellectualism of, 377
 - beginning teachers and, 106
 - descriptive information from, 212
 - evaluative role of, 198-199, 207-208
 - professionalization of teaching and, 366
 - support function of, 362
 - teacher selection and, 376
 - training for, 408-409
 - See also specific types of administrators*
- Affirmative action
- considerations of worth and, 79
 - interviews and, 93-94
 - job-relevant characteristics and, 82
 - teacher selection and, 79, 81-82
 - See also Discrimination, Minority teachers*
- Age, discrimination and, 84
- Albemarle v. Moody*, 349
- Alternate route programs, 66
- American Federation of Teachers, 117, 375
- Analogies, teaching with, 279
- Angoff's method, certification test standards and, 304-306
- Anxiety, evaluation and, 40
- Appearance, teacher selection and, 88, 95
- Appointment blanks, *see* Rating scales
- Architecture, licensed practice in, 64, 283
- Assessment, *see* Evaluation
- Assessment centers, performance exercises in, 281-283
- Assigned work, *see* Classroom assignments
- Attendance, bonus pay for, 231-238
- Attestations, in portfolios, 247
- Attitudes, adult motivation and, 123
- Audiotape
- performance exercises and, 284-285
 - self-assessment and, 218-220
- Authority
- allocating, 31
 - bureaucratic democracy and, 360-368
 - communitarian democracy and, 368-371
 - evaluation and, 37, 358-371
 - expertise as justification for, 362
 - knowledge base and, 362, 365, 367
 - of school board, 363-364
 - sources of, 30
 - See also Governance issues*
- Autonomy, respect for, 358
- Bachelor's degree, graduates with, 47-48
- See also Education programs*
- Basic skills knowledge
- adaptation of, 140
 - career development and, 139-140

- state requirements, 379
- Beginning teachers
 - adult interactions and, 106
 - assistance for, 104-114, 380-382
 - certification test development and, 298, 300-301
 - competencies and, 105-106
 - costs for assistance and, 113-114
 - dropout problem, 104
 - evaluation and, 23, 107-108, 110-114, 142, 380-382
 - professional needs, 105-106
 - programs for, 18
 - socialization and, 106-107
 - successful induction of, 23-24
 - teacher effectiveness and, 105
- Behavior, of teachers
 - category system and, 180
 - classroom observation and, 178
 - context dependent, 184
 - self-assessment of, 218
 - sign system and, 180
 - videotape/audiotape and, 219-220
- Behavior checklists, 177, 180-181
 - classroom observation and, 179-180
 - low inference, 186
 - performance exercises and, 287
 - personal factors and, 185
 - self-evaluation and, 220
 - teaching conceptions and, 185
- Benefit maximization principle, 356
- Bias studies, licensure tests and, 270
- Board of Fort Madison v. Youel*, 344
- Bonuses
 - attendance and, 231-238
 - merit pay and, 230-231
 - principals and, 237
 - teaching minorities and, 231
- Boundary strengthening processes, 42
- Boyce Card, 193-196
- Brag sheets, 221
- Bureaucratic democracy, professionalism and, 360-368, 371
- Business community, evaluation effects on, 43
- California Bar Examination, 283
- Career development, 23-24
 - developmental standards and, 138-141
 - evaluation principles for, 141-143
 - professional development and, 117
 - reflection time and, 168
- Career ladders, 18, 24
 - competence and, 385-386
 - competitive standards and, 137
 - evaluation for, 37, 133, 135-145
 - professional development and, 118
- Career professional teachers, 74
- Carnegie Corporation, 380
- Category system, classroom observation and, 179-180, 188
- Certification, 22, 62-74, 281
 - competency tests and, 18
 - defined, 62-63
 - education programs and, 341
 - evaluation and, 36
 - formal tests for, 28
 - governance issues and, 378-380
 - laws effecting, 30, 341-342, 346-348
 - licensure vs., 281
 - national standards and, 72
 - oral examinations and, 283
 - performance exercises for, 281
 - professional knowledge and, 69
 - professional education courses and, 52
 - semibroad, 68
 - teacher shortages and, 47-48
 - See also* Licensure
- Certification test standards, 295-320
 - adjustment for sampling error, 317
 - Angoff's method, 304-306
 - choice of procedures, 304-311
 - contrasting-groups method, 309-311
 - cut scores and, 295-320
 - data analysis, 312-313
 - data collection, 312
 - documentation and reporting, 319
 - Ebel's method, 307-308
 - errors of measurement and, 315-317
 - experts and, 297-299
 - Jaeger's method, 308-309
 - legal issues and, 347
 - multiple methods and, 318-319
 - Nedelsky's method, 306-307
 - number of judges and, 301-303
 - policy makers and, 297-299
 - scientific basis of, 297
 - selecting judges and, 299-301
 - test results and, 314
 - training judges and, 311-312
 - See also* Licensure tests
- Challenge, teacher responsiveness to, 122-123
- Change, openness to, 121
- Character references, teacher selection and, 80

- Charm, interviews and, 93, 95
- Checklists, *see* Behavior checklists; Rating scales
- Citizenship, 251
- Civil Rights Act (1964), 347
- Classroom assignments
portfolios and, 244
student involvement in, 180
- Classroom documents, growth-oriented
evaluation and, 126
- Classroom management, 87, 206, 226
- Classroom observation, 26, 175-189
beginning teachers and, 111-112
choice of evaluator, 186-187
closed systems, 180-181
conceptions of teaching and, 27
conference after, 186
dismissal records and, 154-155
duration of, 183-184
economic aspects of, 396-397
evaluative inferences and, 181-183
expert teachers and, 222
frequency of, 200-201
generalizability and, 177, 182-185
growth-oriented evaluation and, 125
high inference, 181
job analysis and, 264
low inference, 181
number of times, 183-185
objective outsiders and, 222-223
observer training and, 188
open systems, 179-180
performance assessment and, 71-72
personnel decisions and, 177
portfolios and, 245
procedural considerations, 186
rating scales and, 200-201
reliability and, 177, 181-184
role of, 188-189
sampling occasions, 183-185
systems of, 179-181
Teacher-Evaluation Cycle and, 166-167
teacher selection and, 91, 97
uses of, 178-179
validity and, 177, 182-185
- Classroom Observations Keyed for Effectiveness Research, 180
- Classroom size
beginning teachers and, 109
collective bargaining and, 66
teacher supply and, 65-67
- Classroom track record
beginning teachers and, 104
job applicants and, 91
- Client benefit principle, 358
- Clinical supervision, models of, 36
- Cognitive ability, teacher's
performance and, 327-328
selection interviews and, 377
- Cognitive development, principles of, 205
- Collective bargaining
classroom size and, 66
evaluation and, 129
- Colleges
admissions requirements, 49
freshmen enrollments, 47
grading systems of, 52-56
See also Education programs
- College degrees, requirements for, 46-47
- Collegial products, in portfolios, 247
- Collegial relations, self-evaluation and, 365
- Colloquium sessions, 221
- Communication skills
licensure tests and, 280
process-product research and, 205
- Communitarian democracy, 30, 368-371
- Community
affirmative action and, 82
building a sense of, 204
teacher involvement with, 70, 178
teacher portfolios and, 250
teacher selection considerations and, 83-84
- Compensation systems, *see* Salaries
- Compensatory model, of data combination, 324-330
- Competence
career development and, 139-140
categories of, 67-69
certification and, 62
classroom observation and, 178
competencies and, 261
construct validation and, 273-274
defined, 261
generic, 176
growth in, 24
licensure tests and, 257-275
minimum, 265, 305
performance vs., 261-262
standard setting and, 67-69
- Competencies
beginning teachers and, 105-106
behaviorally oriented, 71
competence and, 261
defined, 261

- wording on licensure tests and, 265-266
- Competency tests, 261
- Complaint-driven evaluation, 148-149
- Con-artists, interviews and, 93, 95
- Connecticut Teacher Assessment Program, 281
- Consultants
 - resources for, 409
 - self-assessment and, 223
 - teacher selection and, 77
- Continuing certification, 51-56, 299
- Continuing education
 - Personnel Evaluation Standards* and, 418
 - portfolios and, 244
 - professional associations and, 385
 - professional development and, 118
 - teacher selection and, 91-92
- Contrasting-groups method, certification test standards and, 309-311
- Course development, career ladders and, 135
- Courts, *see* Legal issues
- Credentialing agencies, 22
- Credentials
 - categories of competence and, 67-69
 - changes over time, 395
 - conceptions of quality and, 69-72
 - duration of, 68
 - emergency, 66
 - grade levels covered by, 68
 - knowledge base and, 69
 - political factors in, 68-69
 - See also* Certification; Licensure
- Criticism, openness to, 121
- Curriculum development, teacher responsibility for, 70
- Curriculum materials, *see* Instructional materials
- Data combination, evaluation and, 29, 322-333
 - choice of method, 330-332
 - clinical vs. statistical methods, 324-330
 - compensatory model, 324-330
 - conjunctive model, 324-330
 - disjunctive model, 324
- Decision making
 - bureaucratic democracy and, 362
 - combining evaluation data for, 29
 - decentralization of, 370-371
 - democracy and, 368-371
 - integration and, 140
 - multiple data sources and, 322
 - role-function combinations and, 139
 - sequential, 329-330
 - teacher participation in, 374
- Democracy, bureaucratic, 360-368
- Demotions, desegregated schools and, 350
- Descriptive scales, 27, 191
 - flexible, 209-211
 - peer observation and, 211-212
 - teacher development of, 211
- Desegregated schools, 350
- Developmental standards
 - career development and, 138-141
 - implementing, 143-144
 - See also* Improvement, performance; Professional development
- Discipline, beginning teachers and, 109
- Discrimination, 30
 - federal law and, 346-350
 - legal issues in evaluation and, 338-339
 - See also* Affirmative action; Minority teachers
- Dismissal, 25, 30
 - control mechanisms and, 42
 - credible evidence and, 154-155
 - desegregated schools and, 350
 - due process and, 348
 - economic issues and, 400
 - evaluation for, 37, 152-156
 - free speech rights and, 348-349
 - political pressure and, 83
 - probationary teachers and, 342-343, 348-349
 - procedural due process and, 155-156
 - rating scales and, 198
 - student achievement and, 337-338
 - student well-being and, 166
 - teacher shortages and, 150
 - tenure and, 343-344, 383
- Documentary history, performance exercises and, 286
- Document review, licensure tests and, 265
- Dossiers, assessment data in, 112-113
 - See also* Portfolios
- Due process, 30, 357
 - dismissals and, 155-156, 348
 - tenure hearings and, 343-344
- Duties statement, teacher selection and, 76
- Dyslexia, 69
- Ebel's method, certification test standards and, 307-308
- Economic issues, in evaluation, 390-401

- classroom observation and, 396-397
- economic efficiency and, 391-392
- measurement of student learning, 397-398
- previous training and experience and, 394-395
- teacher tests and, 395-396
- Educational reform movement, 17, 136-138, 201-203, 350
- Educational Testing Service, 71
- Education practices, landscapes of, 170
- Education programs/schools
 - admissions, 18, 48
 - applicant test standards and, 295-320
 - certification and, 341
 - continuation requirements, 51-56
 - criticism of, 280-281
 - declining enrollment, 376
 - entry tests and, 298
 - grading standards, 52-56
 - licensure test scores and, 274
 - performance tests and, 280-281
 - preservice evaluation and, 375-376
 - professionalism and, 374
 - education requirement and, 46-47
 - state approval of, 379
- Education thought, mindscapes of, 170
- Electives, marginal teachers and, 149
- Elementary schools
 - basic skills progress and, 232
 - career teachers in, 49-50
 - horizontal specialties and, 67
 - pupil-progress criterion, 233
- Employment tests
 - job analysis and, 266
 - legal issues and, 339
 - purposes of, 258
 - scores on, 274
- Engineering, unlicensed practitioners in, 64
- Enrollment numbers, tenure decisions and, 150-151
- Enthusiasm
 - beginning teachers and, 104
 - teacher selection and, 87
- Equal Employment Opportunity Commission, 347
- Equal respect principle, 356-357
- Equal treatment, expectations of, 37
- Essay examinations, student achievement and, 375
- Ethics, in evaluation, 30, 356-372
 - benefit maximization and, 356
 - bureaucratic democracy and, 360-368
 - communitarian democracy and, 368-371
 - equal respect and, 356-357
- Evaluation
 - accountability and, 411
 - authority over, 358-371
 - beginning teachers and, 104, 107-108, 110-114, 380-382
 - bill of rights for, 358-359
 - bureaucratic democracy and, 360-368
 - career ladders and, 133, 135-145
 - collective bargaining and, 129
 - college level, 52
 - communitarian democracy and, 368-371
 - competency-based, 71-72
 - complaint-driven approach to, 148-149
 - consistency in, 40, 42
 - contemporary problem of, 201-205
 - costs of, 113-114, 391-392, 395-396, 399
 - credibility of, 124
 - culture for, 404-410
 - data combination and, 322-333
 - developmental vs. uniform criteria, 142
 - dismissal and, 152-156
 - economic aspects of, 31, 390-401
 - effects of, 20-21, 35, 37-44
 - ethics and, 356-372
 - external pressures for, 200-203
 - federal law and, 346-350
 - formative vs. summative, 143
 - governance issues in, 371-388
 - historical perspective, 191-200
 - implementation issues, 19-20, 31-32, 403-410
 - incentives and, 392-393
 - individual level effects, 38-41
 - input oriented, 391
 - legal issues and, 337-351
 - merit pay and, 133-135, 138-145
 - methods of, 25-29
 - multiple data sources, 110-112, 322-333
 - objectives of, 204-205
 - on-the-job performance and, 26-28
 - organization level effects, 38, 41-43
 - outcomes of, 19
 - output oriented, 391
 - parent surveys and, 111
 - performance exercises and, 278-292
 - Personnel Evaluation Standards* and, 416-428
 - portfolios and, 241-255
 - preservice, 46-59, 375-376

- previous training and employment and, 394-395
- production theory and, 391-392
- professional development and, 116-131, 323
- professional standards in, 365-366
- pupil surveys and, 111
- purposes of, 21-25, 35-37, 116, 177, 158-159, 323-324, 374-375
- requirements of, 420-425
- research on, 71, 203-204
- resources for, 409-410
- school improvement and, 25, 158-170
- selection decisions and, 77
- self-evaluation and, 191-213, 216-227
- simulations and, 278-292
- skill-based vs. performance-based, 141-142
- stakeholder involvement in, 406-407
- state requirements, 340-346, 379
- student achievement and, 110, 229-239
- subjective vs. objective, 142-143
- superintendent's commitment to, 407-408, 410-411
- sustaining, 410-413
- taxpayer welfare and, 399
- teacher control of, 112-113, 191-213
- Teacher-Evaluation Cycle, 166-169
- teacher shortages and, 147-148
- tenure and, 150-152
- Evaluative inferences
 - classroom observation and, 181-183, 186
 - licensure tests and, 260-263
- Evaluators
 - classroom observation and, 176, 186-187
 - credibility of, 423
 - growth facilitation and, 123-125
 - knowledge base of, 207
 - rating scales and, 198
 - teachers as, 168, 187
 - training for, 181, 188
 - trust in, 404
- Evidentiary hearings, dismissals and, 343-344
- Excellent teachers, 23
 - designing licensure tests and, 266-267
 - formative evaluation and, 217
 - identification of, 204
 - performance exercises and, 290-291
- Experts
 - certification test standards and, 297-299
 - licensure test development and, 263
 - subject matter competence and, 367-368
 - teacher improvement and, 384
- Expert teachers, 222-223, 288
- Extracurricular demands, teacher selection and, 377
- Fairness, norms regarding, 42-43
- False-negative errors, certification tests and, 316
- False-positive errors, certification tests and, 316-317
- Feasibility standards, *Personnel Evaluation Standards* and, 423-424
- Federal law, evaluation and, 346-350
- Feedback
 - beginning teachers and, 107
 - growth-oriented evaluation and, 124, 126-127
 - improvement and, 38
 - nonjudgmental, 217
 - systematic, 118-119
 - timing of, 48, 126
 - videotape/audiotape, 218-220
- Field tests, of licensure tests, 269-270
- Flanders Interaction Analysis System, 180
- Florida Performance Measurement System, 176-177, 180-181
- Follow-up activities, performance improvement and, 38
- Formative evaluation, 216-217
 - effectiveness of, 27
 - self-assessment and, 217-227
 - See also Professional development
- Free choice, 365
- Free speech rights, 348-349
- Future orientation, continued growth and, 122
- General ability, performance and, 326-328
- General checklists, 197
- Goal consensus, evaluation effects on, 41
- Goal setting, 25
 - performance improvement and, 124
 - school improvement and, 160-161
 - self-comparison and, 223-224
- Governance issues, 371-388
 - advancement and, 385-386
 - beginning teachers and, 380-382
 - certification and, 378-380
 - improvement and, 384-385
 - licensure and, 378-380
 - merit pay and, 385-386

- preservice evaluation and, 376
- professional development and, 384-385
- retention and, 382-383
- school improvement and, 387
- teacher selection and, 376-378
- tenure and, 382-383
- Grade level, teaching variations and, 183
- Grade-point average, teacher's
 - calculation of, 53-54
 - success in teaching and, 54-55
 - teacher selection and, 80
- Grading standards
 - education schools, 52-56
 - practice teaching and, 56-57
- Griggs v. Duke Power Company*, 339
- Group discussion, licensure tests and, 265
- Growth
 - conditions associated with, 120-121
 - evaluation and, 119
 - facilitators of, 123-125
 - risk taking and, 128
 - See also Professional development
- Guided entries, in portfolios, 247
- Higher-order thinking, 206
- Hiring, see Teacher selection
- Homebase room, for beginning teachers, 109
- Home-study teachers, marginal performance and, 149
- Humaneness rights, 357
- Human inputs, costs of, 399
- Identity diffusion, 122
- Improvement, performance, 23-25
 - classroom observation and, 177
 - developmental standards and, 138
 - environmental level and, 43-44
 - evaluation and, 405-406
 - governance issues and, 384-385
 - rating scales and, 197, 198-199, 212
 - self-assessment and, 216-227
 - tenured evaluation and, 383
 - See also Professional development; School improvement
- Individual level effects, of evaluation, 36, 38-41
- Induction programs, 104, 113-114
- Information, economics of, 392
- Inservice education, 109, 222
- Institutional forms, political values and, 30
- Instructional approach
 - behavior variables and, 183
 - classroom observation and, 178
 - teacher-directed, 177
- Instructional materials
 - portfolios and, 244
 - strength of, 279
- Instructional vignettes, 284-285
- Intellectual ability, teacher selection and, 87
- Intellectual exchange, need for, 208
- Internships, 72, 325
 - funding for, 381
 - professional development school and, 374
- Interviews
 - affirmative action and, 93-94
 - classroom observation and, 189
 - problems with, 92-95
 - selection decisions and, 77, 92-97, 377
 - self-assessment and, 223
 - telephone, 95-97
- Isolation, of teachers, 39, 107, 125, 248
- Items-to-observe checklist, 197
- Jaeger's method, certification test standards and, 308-309
- Job analysis
 - classroom observation and, 264
 - licensure tests and, 264-267
- Job description
 - self-comparison to, 223-224
 - teacher selection and, 76, 81, 98-99
- Job enrichment, professional development and, 118
- Job skills, assessment centers and, 282
- Joint Committee on Standards for Educational Evaluation, 32, 416-428
- Jungian psychology, self-assessment and, 221
- Junior high teachers, credential for, 68
- Knowledge base, authority and, 362, 365, 367
- Kudasik v. Board*, 345
- Labor market
 - demand and, 47-48
 - education standards and, 51
 - monopolies of competence in, 64
 - teacher supply and, 22
- Leadership skills, 140, 386
- Learning disabled, low achievers vs., 69
- Learning styles, self-assessment of, 221

- Legal issues, 30
 bureaucratic democracy and, 360-368
 certification and, 341-342, 346-348
 courts' enhanced role and, 337-340
 dismissal of probationary teachers and, 342-343
 dismissal of tenured teachers and, 343-344
 evaluation and, 337-351
 federal law and, 346-350
 free speech rights, 348-349
 liberty and property interests, 348
 performance evaluation laws and, 344-346
 state law and, 340-346
- Lesson planning, 185
 beginning teachers and, 109
 classroom observation and, 178
 growth-oriented evaluation and, 126
 performance exercises and, 287
 portfolios and, 244, 250
 time for, 208
- Lesson type, teaching variations and, 183
- Letters of recommendation, 80
- Liberty interests, dismissals and, 348
- Licensure, 22, 62-74
 certification vs., 281
 compensatory model and, 330-331
 conjunctive model and, 330
 defined, 62-63
 evaluation for, 323
 governance issues and, 378-380
 performance exercises for, 281
 professionalism and, 374
 professional knowledge and, 69
 See also Certification; Licensure tests
- Licensure tests, 28, 257-275
 advantages of, 259
 bias studies and, 270
 cut scores, 64-65, 262, 272
 development procedures, 263-269
 domain specifications, 268
 field test of, 269-270
 inferences from, 260-263
 item writing and review, 269
 job analysis and, 264-267
 limitations of, 259-260, 280
 outline and objectives, 263-264
 purposes of, 258-259
 reliability of, 271
 reporting results, 274
 small-scale pilot of, 269
 validity and, 262-263, 271-274, 280
- Literacy
 licensure tests and, 280
 selection decisions and, 97
- Low achievers, learning disabled vs., 69
- Mantel-Haenszel technique, 270
- Master-teacher plans, *see* Career ladders
- Mathematics teachers, 206
 bonuses for, 231
 performance exercises and, 283-287
- Math scores, education majors and, 49-50
- Mean scores, 313
- Measurement error, 315-317
- Median scores, 313
- Medicine
 labor supply, 64
 specialization in, 67
- Mentors
 beginning teachers and, 109
 career ladders and, 135
 funding for, 381
 portfolios and, 247
- Merit
 defined, 23
 distinguishing worth from, 84-85
 job-specific criteria for, 98
 professional development and, 118
 teacher selection and, 78-81, 84-86
 tenure and, 151-152
- Merit pay, 24
 bonuses and, 230-231
 competitive standards and, 137
 economic issues and, 399
 evaluation for, 133-135, 138-145, 323
 governance issues and, 385-386
 procedures for, 18
 rating scales and, 198
 student achievement and, 230-238
 subjectivity and, 134
- Merit rating, 198
- Metaphors, teaching with, 279
- Microteaching, 219-220, 279
- Middle school teachers, credentials for, 68
- Minority students
 bonuses for teaching, 231
 marginal teachers and, 148-150
 tenure decisions and, 151, 156
- Minority teachers
 certification tests and, 30, 347
 legal issues and, 338-339
 licensure test development and, 263
 National Teachers' Examination and, 347

- See also* Affirmative action
- Mission statement, teacher selection and, 76, 81
- Modeling, self-assessment and, 222
- Motivation, teacher's, 21
evaluation effects on, 39-40
grading systems and, 53
growth facilitators and, 124-125
role of attitude in, 123
system-wide planning and, 162
- Motivation, students
devices for teaching and, 279
intrinsic, 211
- Mt. Healthy v. Doyle*, 348-349
- Multiple-choice exams, licensure and, 258-260
- Narrative records, classroom observation and, 179-180
- National Board for Professional Teaching Standards, 22, 28, 72, 380, 385
- National Board of Medical Examiners, 283
- National Commission on Excellence in Education, 201
- National Counsel for Accreditation of Teacher Education, 50, 379
- National Education Association, 117, 374-375
- National surveys, teachers judged by, 203
- National Teacher Examination, 18, 258, 298
bias and, 270
false negatives and, 65
minority applicants and, 347
professional knowledge and, 71
- Nedelsky's method, certification test standards and, 306-307
- Needs assessment, teacher selection and, 76-77, 81
- News media, teachers judged by, 201-203
- Nonverbal cues, self-assessment and, 218
- Normal schools, 48
- Numeracy, licensure tests and, 280
- Nursing, labor supply and, 64
- Observation, *see* Classroom observation
- Observation forms, self-evaluation and, 220
- Observation Schedule and Record, 180
- Oklahoma Teacher Reform Act (1980), 187
- Oral examinations, certification and, 283
- Organizational development, professional development and, 117
- Organizational level effects, of evaluation, 38, 41-43
- Orientation sessions, 108
- Overall value incentive system, 393
- Parent(s)
complaints from, 148-149
evaluation effects on, 43-44, 407
- Parent interactions, 70
beginning teachers and, 106
classroom observation and, 178
intellectual teachers and, 377
portfolios and, 250
- Parent ratings, 369-371
beginning teachers and, 111
descriptive, 212
- Pedagogical knowledge
content-specific, 279, 288-289
documentary evidence and, 286
licensure tests and, 264, 267-268
performance exercises and, 284-285
profession of teaching and, 367-368
state requirements, 379
- Peer review
beginning teachers and, 111
classroom observation and, 187, 211-212
professional development and, 125, 129
professionalism and, 362
self-assessment and, 219, 225
team evaluation and, 187
- Performance
cognitive ability and, 327-328
competence vs., 261-262
consistency of, 183
defined, 261
evaluation and, 26, 36
general ability and, 326-328
impact of information on, 38-41
incentives for, 31, 392-393
job knowledge and, 262-263
norms for, 41
peak, 122
stability of, 183
teacher effectiveness vs., 261-262
time-pattern of, 80
See also Improvement, performance
- Performance assessment, *see* Evaluation
- Performance exercises, 278-292
administration of, 289
alternate forms for, 290

- assessment centers and, 281-283
- documentary history and, 286
- examples of, 283-288
- instructional vignettes and, 284-285
- knowledge base and, 288-289
- scoring, 287-288
- topic sequencing and, 285-286
- validity and reliability of, 290-292
- Performance improvement, *see* Improvement, performance
- Performance objectives
 - classroom observation conferences and, 167
 - school improvement and, 160-161, 163-164
 - team approach to, 163-164
- Performance standards, *see* Standards
- Personal development, 21
 - professional development and, 117, 130
 - self-assessment and, 218
- Personal disorders, poor performance and, 154
- Personality, 200-201
 - interviews and, 94-95
 - Jung's theory and, 221
 - teacher selection and, 88
- Personal life
 - beginning teachers and, 106
 - teacher selection and, 88
- Personnel decisions, 217
 - classroom observation and, 177
 - evaluation and, 36
 - legal issues and, 337-351
 - portfolios and, 243-244, 248
 - subjective, 208
 - See also* Dismissal; Teacher selection; Tenure
- Personnel Evaluation Standards*, 32, 416-428
 - accuracy standards and, 424
 - application of, 425-427
 - background of, 416-418
 - development of, 418-419
 - feasibility standards and, 423-424
 - guiding definitions for, 419-420
 - proprietary standards and, 420-422
 - utility standards and, 422-423
- Personnel file, 168, 242
- Planning, *see* Lesson planning
- Policy making, 20
 - bureaucratic democracy and, 360
 - certification test standards and, 295, 297-299
 - licensure test scores and, 274
 - new reform movement and, 201-203
 - professional development and, 127-130
 - teacher evaluation and, 18
- Political pressure, standards boards and, 380
- Portfolios, 27-28
 - activities shown in, 252-253
 - contents of, 248-253
 - organization of, 246-248
 - purposes for, 243-245
 - structure for, 249-253
 - teacher, 241-255
- Post-observation conference, 225
- Practice teaching
 - classroom track record and, 91
 - grading standards and, 56-57
 - preservice evaluation and, 56-57
 - requirement of, 48
- Preobservation conference, 224
- Preparation
 - Personnel Evaluation Standards* and, 418
 - portfolios and, 250
 - rating scales and, 197
 - time for, 208
 - See also* Lesson planning
- Preservice evaluation, 22, 46-59
 - education schools and, 375-376
 - nature and importance of, 46-48
 - practice teaching and, 56-57
- Principals
 - assessment centers and, 282
 - bonus pay and, 237
 - classroom observation and, 179, 186-187
 - conflicting roles for, 107, 110
 - growth facilitation and, 125
 - school improvement teams and, 162
 - teacher selection and, 77
- Privacy rights, 357, 422
- Probationary teachers
 - dismissal of, 342-343, 348-349
 - federal law and, 348-349
 - state law and, 342-343
 - See also* Beginning teachers
- Problem solving skills, teaching, 206, 398
- Process-product research, 205-207
- Proctored entries, in portfolios, 247
- Production theory, 31, 391-392
- Professional development, 21, 23
 - accountability and, 127-130
 - beginning teachers and, 105-106

- continuing, 118
- defined, 117
- evaluation and, 38-39, 116-131, 159, 323
- failure and, 122, 128
- formative evaluation and, 217
- governance issues and, 384-385
- individual characteristics and, 121-125
- induction process and, 114
- justification for, 117-118
- mechanisms for, 118
- performance objectives for, 160
- policy making and, 127-130
- portfolios and, 243-245, 253
- rating scales and, 198
- research on, 119-121, 130-131
- resources and, 129, 409-410
- school districts and, 384-385
- self-assessment and, 226
- system characteristics and, 125-127
- standards and, 138
- Teacher-Evaluation Cycle and, 168
- See also* Growth
- Professional associations, continuing education and, 385
- Professional education courses, requirements for, 51
- Professionalism, 17, 30
 - accountability and, 365-366
 - bureaucratic democracy and, 360-368, 371
 - democracy and, 359-371
 - institutionalization of, 363-364
 - nondemocratic values and, 362-363
 - reform aims of, 374-375
 - state involvement and, 378-380
- Professional Knowledge Test, 65
- Professional standards boards, 374, 378-379
- Professors, licensure test development and, 263, 298
- Promotion
 - career ladders and, 135-136
 - compensatory model score and, 328
 - developmental standards and, 139
 - governance issues and, 385-386
 - merit and, 79
 - professional development and, 118
 - quotas and, 143-144
 - rating scales and, 198
 - structural obstacles to, 135
- Property interests, dismissals and, 348
- Proprietary standards, 32, 420-422
- Psychological characteristics, measurement of, 273
- Psychopaths, interviews and, 93, 95
- Public opinion, teachers judged by, 203
- Public perspicuity rights, 357
- Pupils, *see* Students
- Quality
 - merit and, 85
 - standard setting and, 69-72
- Quality scales, 193
- Quantity of work, merit and, 85
- Questionnaires, self-assessment and, 223
- Quotas
 - bonus pay and, 231
 - competitive standards and, 137-138
 - promotion and, 143-144
- Rating systems, 27
 - classroom observation and, 179-181
 - 200-201
 - conceptions of teaching and, 27
 - contemporary problems with, 201-205
 - content of, 197-198
 - difficulties with, 199
 - descriptive, 209-213
 - future role of, 207-209
 - historical perspective, 191-200
 - historical problems with, 200-201
 - licensure tests and, 262
 - qualitative, 193
 - quantitative, 193
 - self-evaluation and, 220
 - use and consequences of, 198-199
- Reading competency scores, teacher's, 49-50
- Reading concepts, teaching, 206
- Reassignment, of teachers, disadvantaged students and, 156, 350
- Recruitment, market forces and, 47
- Reflection, time for
 - career planning and, 168
 - professional development and, 245
- Reflection-in-practice, 131
- Registered nurses, 64
- Release time, professional development and, 118, 129
- Religious schools, 81, 83
- Representative democracy, 368-371
- Research
 - evaluation systems based on, 36
 - overgeneralization of findings, 206-207
 - process-product, 205-207

- teacher effectiveness and, 192-193, 207
- Resignations, induced, 149-150
- Retention
 - governance issues and, 382-383
 - market forces and, 47
- Retirement, early, 149-150
- Rewards, allocation of, 28
- Risk taking, positive orientation to, 121
- Rural schools, 76, 81
- St. Louis Teacher's Union v. Board of Education*, 338
- Salaries
 - politics and, 64
 - teacher supply and, 63-64
 - See also Bonuses; Merit pay
- SAT scores, teacher training institutions and, 49-50
- Scheelhaase v. Woodbury*, 337-338
- School boards
 - authority of, 363-364
 - beginning teachers and, 106
 - performance improvement and, 44
 - teacher selection and, 77
- School districts
 - beginning teacher assistance and, 380-382
 - as bureaucratic democracy, 360-368
 - evaluation effects and, 43, 407
 - management information systems and, 378
 - performance expectations and, 39
 - professional development and, 384-385
 - reform movement and, 202-203
 - teacher selection and, 376-378
- School improvement, 25
 - evaluation for, 158-170
 - external pressures for, 201-202
 - goal-setting models and, 160-161
 - governance issues and, 387
 - integrating with staff development, 164-165
 - performance objectives and, 163-164
 - portfolios and, 243
 - setting focus for, 161-162
- Science, bonuses for teaching, 231
- Science scores, teacher training institutions and, 49-50
- Score cards, see Rating scales
- Secondary school teachers
 - basic skills progress and, 232
 - horizontal specialties, 67-68
 - improvement objectives and, 162
 - pupil-progress criterion, 233
 - subject knowledge and, 70, 87
- Selection, see Teacher selection
- Self-assessment, 27, 191-213, 216-227
 - collegial relations and, 365
 - defined, 218
 - hybrid systems, 219, 224-226
 - peer review and, 219
 - problems with, 226-227
 - self-improvement and, 27
 - techniques of, 219-224
- Self-esteem, beginning teachers and, 106-107
- Senior teachers, evaluation criteria for, 142
- Showcase lessons, 184-185
- Sick leave, cost of substitutes and, 231-232
- Sign systems, classroom observation and, 180
- Simulations, 29, 185, 397. See also Performance exercises
- Singleton v. Jackson*, 350
- Skill-based evaluation, 141-142
- Small-scale pilot, of licensure tests, 269
- Socialization
 - beginning teachers and, 104, 106-107
 - evaluation and, 24
 - of students, 208
- Social reform, court involvement in, 339
- Social studies, performance exercises and, 283, 286-287
- Software, cost-effectiveness of, 399
- Speaking skills, telephone interviews and, 95
- Special education teachers, categories of, 68
- Specialization
 - credential system and, 67
 - horizontal, 67
 - interest group politics and, 68-69
 - vertical, 67
- Spelling test, teacher selection and, 97
- Standard error of measurement, 316-317
- Standards, 29, 32, 365-366
 - academic, 377
 - certification and, 281, 295-320
 - clarity of, 125
 - competitive, 24-25, 137-138
 - developmental, 24-25, 125, 138, 143-144
 - due process and, 357
 - education reform and, 136-138

- federal law and, 349-350
- flexibility in, 64-65
- informing teachers of, 186
- licensing boards and, 378-380
- minimum, 136-137
- rating scales and, 198
- self-assessment and, 223-224
- See also Certification test standards
- Standards for Educational and Psychological Testing*, 296, 347
- Standards for Evaluation of Educational Personnel*, 322
- Standards for Evaluations of Educational Programs, Projects, and Materials*, 416
- State governments, 31, 43
 - evaluation policies and, 18
 - legal issues in evaluation and, 340-346
 - teacher supply and, 22
- Stress, evaluation and, 40
- Student(s)
 - evaluation of, 250
 - as individuals, 251-252
 - interaction with, 193
 - involvement in evaluation and, 369-371
 - labor market experiences and, 395
- Student achievement
 - dismissal and, 337-338
 - economic issues and, 397-398
 - essay examinations and, 375
 - evaluation and, 27, 110, 229-239
 - growth-oriented evaluation and, 125-126
 - growth standards and, 231-233
 - licensure tests and, 262
 - merit pay and, 230-238
 - portfolios and, 253
 - process-product research and, 205
 - teaching effectiveness and, 176-177
- Student interviews, self-assessment and, 223
- Student needs, system-wide planning and, 161
- Student ratings
 - beginning teachers and, 111
 - descriptive, 212
 - teacher effectiveness and, 192
 - teacher selection and, 91
- Student teachers, 375
- Study materials, self-assessment and, 221
- Subject knowledge, 176
 - beginning teachers and, 112
 - changing over time, 91, 118
 - classroom observation and, 178, 187
 - credential system and, 69-71
 - job performance and, 262-263
 - licensure tests and, 264
 - performance exercises and, 279, 284
 - portfolios and, 251
 - professionalism and, 367-368
 - state requirements, 379
 - teacher selection and, 87, 91-92
- Subject matter, teaching variations and, 183
- Substitutes
 - costs of, 231-232
 - credentials for, 66
 - marginal teachers and, 149
- Superintendents
 - classroom observation and, 186-187
 - commitment to evaluation, 407-408, 410-411
 - teacher selection and, 77
- Supervisors, 154
 - classroom observation and, 179, 186-187
 - growth facilitation and, 125
 - licensure test development and, 263
 - portfolios and, 245, 253
 - professional development and, 118, 129
 - rating scales and, 198
 - responsibilities of, 198
 - student teachers and, 375
- Supplies, beginning teachers and, 105, 109
- Taxpayer welfare, evaluation and, 399
- Teacher(s)
 - activities of, 250-251
 - as evaluators, 168, 187
 - attrition rates, 23
 - dissatisfaction of, 208-209
 - evaluation control and, 112-113
 - functions of, 139
 - incompetent, 257
 - isolation of, 39, 107, 125, 248
 - licensure test development and, 263
 - misassignment of, 66
 - occupational structure, 74
 - portfolios of, 241-255
 - public ambivalence about, 203
 - qualities to look for, 86-87
 - roles of, 139
 - selection of, 76-100
 - self-assessment, 27
 - study groups of, 211-212
 - teacher selection and, 77

- See also Teaching*
- Teacher Assessment Project, 281, 283, 287
- Teacher characteristic profile, 197
- Teacher effectiveness
- beginning teachers and, 105
 - consensus on, 175-176
 - context and, 204
 - defined, 261
 - early research on, 192-193
 - grade-point average and, 54-55
 - licensure tests and, 262
 - performance exercises and, 279
 - performance vs., 261-262
 - previous employment and, 394-395
 - practice teaching performance and, 56
 - process-product research and, 207
 - rating scales and, 199-200
 - student achievement and, 176-177
 - student ratings and, 192
- Teacher-Evaluation Cycle, 166-169
- Teacher Evaluation Needs Identification Survey, 120
- Teacher Evaluation Profile Questionnaire, 120-121, 124
- Teacher selection, 22-23, 76-100
- ability and, 87
 - affirmative action and, 79, 81-82
 - classroom observation and, 91, 97
 - compensatory model and, 330
 - data gathering and, 89-97
 - evaluation for, 36, 159, 323
 - false negatives and, 98
 - governance issues and, 376-378
 - initial stages of, 22-23
 - interview and, 77, 92-97
 - job description, 76, 81, 98-99
 - logistics of, 377-378
 - merit and, 78-81, 84-86
 - needs assessment and, 76-77, 81
 - nonclassroom tasks and, 81
 - preferred characteristics and, 400
 - previous employment and, 394
 - procedures for, 76-77
 - school districts and, 376-378
 - subject matter knowledge and, 91-92
 - teaching style and, 88-89
 - telephone interviews and, 95-97
 - verifying application data and, 89-90
 - worth and, 78, 81-86
- Teacher shortages, 341
- dismissal and, 150
 - history of, 47
 - state requirements and, 380
 - tenure and, 147-149
- Teacher supply
- classroom size and, 65-67
 - credential system and, 63-67
 - salaries and, 63-64
 - test scores and, 64-65
- Teacher traits, *see* Rating scales; Traits
- Teaching
- alternate route programs into, 66
 - behavior variability and, 26
 - complexity of, 200
 - context dependent, 184
 - defining, 185
 - functions of, 206
 - knowledge base of, 288-289, 362, 365, 367
 - specialization in, 67-69
 - task clusters and, 250
 - work-place conditions of, 208
- See also Teacher(s)*
- Team activities
- peer review and, 187
 - school improvement and, 162-164
- Team teaching, 222
- Technical knowledge
- of evaluator, 124
 - teacher quality and, 70-71
- Telephone interviews, teacher selection and, 95-97
- Tenure
- compensatory model score and, 328-329
 - dismissal and, 343-344, 383
 - enrollment numbers and, 150-151
 - evaluation and, 25, 36-37, 150-152, 323
 - governance issues and, 382-383
 - merit-based, 151-152
 - Teacher-Evaluation Cycle and, 166-169
 - teacher shortages and, 147-149
- Testing knowledge, teacher selection and, 87, 92
- Threshold-loss function, 271
- Topic sequencing, performance exercises and, 285-286
- Track record, teacher selection and, 79-80
- Training
- administrators and, 408-409
 - career ladders and, 135
 - classroom observation and, 188
 - evaluation based on previous, 394-395
 - for evaluators, 181, 188
 - for growth facilitators, 125
 - inadequate funds for, 207-208
 - professional development and, 129

- standard-setting judges and, 311-312
- Traits, 192
 - assessment centers and, 282
 - licensure tests and, 259-260
- Transfers, 123
 - disadvantaged students and, 155
 - parent complaints and, 148-149
 - professional development and, 118
 - rating scales and, 198
- Travel funds, professional development and, 129
- Trust, formative evaluation and, 216-217
- Tuition reimbursement, 129
- U.S. history, performance exercises and, 283, 286-287
- Utilitarianism, 356
- Utility standards, 32, 422-423
- Verbal cues, self-assessment and, 218
- Videotape, 391
 - costs of, 397
 - performance exercises and, 279, 285
 - self-assessment and, 218-220
 - training observers with, 188
- Visitation programs, beginning teachers and, 109
- Vocational nurses, 64
- Watson v. Fort Worth Bank and Trust*, 339, 349
- Worth
 - affirmative action and, 82
 - defined, 23
 - distinguishing merit from, 84-85
 - institution-specific criteria for, 98
 - sources for standards of, 81
 - teacher selection and, 78-81, 81-86
- Writing project workshops, 123

THE NEW HANDBOOK OF TEACHER EVALUATION

The New Handbook of Teacher Evaluation

Edited by *James H. Gage, University of California, Los Angeles*

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5891-1

Hardcover, 304 pages, \$49.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5892-9

Paperback, 304 pages, \$29.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5893-7

Hardcover, 304 pages, \$49.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5894-5

Paperback, 304 pages, \$29.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5895-3

Hardcover, 304 pages, \$49.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5896-1

Paperback, 304 pages, \$29.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5897-9

Hardcover, 304 pages, \$49.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5898-7

Paperback, 304 pages, \$29.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5899-5

Hardcover, 304 pages, \$49.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5899-5

Paperback, 304 pages, \$29.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5899-5

Hardcover, 304 pages, \$49.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5899-5

Paperback, 304 pages, \$29.95

Published by *Corwin Press, Inc., 2000*

ISBN 0-8059-5899-5

Hardcover, 304 pages, \$49.95

Published by *Corwin Press, Inc., 2000*

